Visual Perception in Simulated Reality

Swafford, Nicholas

Award date:
2016

Awarding institution:
University of Bath

Visual Perception in Simulated Reality

submitted by

Nicholas Teixeira Swafford

for the degree of Master of Philosophy

of the

University of Bath

Department of Computer Science

April 2016

Contents

Abstract

This thesis addresses the use of features of human visual perception to alleviate computation time for real-time computer graphics applications and the initial steps towards the construction of a perceptually informed virtual reality environment. Using a subset of gaze-contingent rendering techniques, named perceptually lossless foveated rendering techniques, real-time rendering systems are able to selectively render at much lower fidelity in a way that the average user is unable to distinguish any difference in quality. This is achieved through the use of an eye tracking device, in order to render a fixed region around the user's point of gaze on a display surface at high quality with the rest of the render (which is in their peripheral field of view) at lower spatial quality.

Although foveated rendering techniques have been explored in the past, it is only more recently that eye tracking and supporting rendering hardware have reached a point which reliably enables development of perceptually lossless (indistinguishable) investigation. Additionally, the resurgence of virtual reality in the commercial sector (along with its demands for rendering quality that exceeds the capability of many modern, commercially-accessible hardware) justifies the study and adoption of these methods.

As such, this thesis presents the work conducted as part of an Engineering Doctorate in Digital Media with the Centre for Digital Entertainment in collaboration with Disney Research on the construction of a commercially accessible, enthusiast level, perceptually augmented virtual reality environment. This includes research towards novel perceptually lossless foveated rendering methods, addressing concerns specific to gaze-contingent methods, some of the engineering efforts towards the construction of the virtual reality environment, and additional related work conducted specifically as part of the collaboration with Disney Research.

Chapter 1

Introduction

1.1 Motivation

The field of graphics rendering is dominated by algorithmic and hardware optimizations that expect a render to be fully appreciated at any single point in time. Despite its huge significance, the user's perception is assumed to be perfect, despite even some of its most obvious flaws. This need has held back the adoption of more physically accurate methods with higher quality results, normally the domain of cinematic rendering, in real-time rendering applications. When interactivity is critical, the primary focus remains on how much can be rendered rather than how little needs to.

Through an extensive study of the human visual system, knowledge of its strengths and weaknesses can be integrated into rendering to alleviate total computational load. Consequently, computationally intensive techniques can be applied selectively, producing results that appear to be of much higher quality where it truly matters. Ideally this should be done in a way that, when presented with a selectively rendered scene and a fully rendered scene, a user would be unable to distinguish the two. In this event, perceptually lossless rendering is achieved.

The development of perceptually lossless models is motivated by two elements. Firstly, it is only recently that the state of graphics hardware and commercial demands have reached a stage that force us to make the next leap forward. The resurgence of commercial interest in Virtual Reality (VR) has prompted industry and academia to quickly find ways of delivering results that differ in quality by factors of two, three, or more. For example, pixel densities that approach the limits of human vision for standard viewing at arm's length (such as tablets or desktop screens) are nowhere near the pixel densities required to achieve the same effect with popular virtual reality hardware such as Head-Mounted Displays (HMDs). Secondly, it is only recently that

enabling technologies have reached a critical point that makes the development of perceptually lossless models seems feasible.

1.2 Contents

This thesis outlines several threads of research and engineering conducted during this project. Most of the work was originally intended to contribute towards the construction of a perceptually-aware virtual reality installation as a doctoral project. As such, this thesis mostly describes some of the initial and intermediate steps taken in this direction. The work discussed in this thesis was originally meant to contribute a doctoral project much larger in scope. Additionally, the work was conducted with an industry partner as part of the doctoral program, namely Disney Research. Given the pull of commercial/industrial interests, the current body of work reads more as a portfolio of contributions rather than a single unified goal, with effort spread across multiple areas.

Background literature is presented in Chapter 2, outlining several perceptual phenomena and relevant works required to understand the motivation and direction of the remaining chapters. The chapter is divided by topics addressing ocular physiology, visual psychophysics, a brief introduction to eye-tracking methods, virtual reality hardware, display technology, a brief introduction to rasterization and ray-based rendering, and finally existing work on gaze-contingent rendering methods. The latter forms an integral part of the optimization work for this project. It is important to note that this thesis is not meant to serve as a record of all research in these topics; instead, it only outlines highly relevant research that directly informed and guided the work.

Chapters 5, 4, & 3 describe most of the work conducted throughout the program. It excludes project work conducted exclusively for and internally within the partner company. These chapters should be read as a portfolio of work addressing the research questions in Section 1.3. As mentioned, this section should be read as a portfolio of relevant work contributing towards the ultimate goal of constructing a perceptually aware virtual environment. The chapter is divided by publications, patents, and general contributions. A general contribution refers to work that was not included in any individual paper or patent to date, but would eventually lead to one or simply be a significantly novel aspect of the perceptually-aware system. Finally, Chapter 6 outlines how all threads of work in this project tie together and outlines future work for the project's continuation.

Some of the work in this thesis has been accepted as two short paper publications in conference proceedings where I am listed as primary author, one published at Virtual

Reality Software and Technology (VRST) in 2014 (Section 5.1) and the other at Conference on Visual Media Production (CVMP) 2015 (Section 4.1). I have also contributed to a short paper for CVMP 2014 (Section 5.2). Additionally, one patent in which I am listed as an inventor has been submitted to the US patent office (Section 3.1). The work in Section 4.2 has recently been submitted and will be awaiting a publication decision.

1.3 Research Problem Overview

The results of this research are of primary interest to the video game and real-time rendering industries, or those industries wishing to breach real-time rates. However, the tight coupling between perception and VR, as well as the huge industrial push for commercially accessible VR hardware and content, cements the need for further perceptual research in the field. Most prior work has focused on applying perceptual rendering methods in a lossy manner. To achieve lossless quality, an in-depth exploration of visual psychophysics is required.

Additionally, only recently have companies started looking into effectively exploiting human visual perception in a commercial setting. This includes the use of eye-trackers for purposes other than active user input. Constructing these systems at a commercial level is novel enough to warrant discussion. What this work sought to answer, therefore, can by summarized by the following questions:

Novel Uses for Perception

How can we further exploit perception in real-time rendering? Looking at existing models, how can these be extended to cover further limitations of the visual system? In which domains does it make most sense to apply these novel techniques?

Lossless Foveation

How can these exploits be achieved in a lossless way? As mentioned prior, the primary academic goal of this work is the development of lossless models, which are not sufficiently addressed in prior literature. How can the methods we develop be refined in a way such that lossless degradation is achieved? Which models provide us with the most computational gain and the least amount of subjective quality loss?

Perception and Hardware

How can we combine existing hardware to exploit these methods?

Given what is available today, how can we construct Virtual Reality Environments (VREs) that exploit human perceptual systems as much as possible? How can the models developed above be successfully integrated into existing and upcoming VR systems? What aspects of VR systems enable or facilitate the implementation of these methods?

Commercial Relevance

How can this technology be used effectively in the commercial domain? Given the current commercial climate and the nature of the host technology itself, VR devices (and specifically rendering for VR) stands to benefit the most from novel developments within perceptual rendering. How can a fully perceptually aware VRE enable further perceptual exploits? How can perceptual information be exploited in VREs beyond the rendering pipeline (such as gaze reactive characters, gaze-dependent information systems, etc.)?

These questions are addressed multiple times throughout my work documented in this thesis. This includes details on efforts in the direction of a cost-effective VRE (Sections 5.4 & 5.3), the development and implementation and evaluation of new and existing foveated rendering models (Section 4.2), their implementation in a commercial game engine (Section 4.1), and cheaper/commercially-accessible alternatives for VR hardware (Sections 5.1 & 5.5) among other work.

Chapter 2

Background

2.1 Visual System Physiology

A complex organ, the eye is the entry point for light to be processed by the brain. Incoming light excites rod and cone cells, sensitive to light intensity and colour respectively, sending these signals via the retinal ganglion cells. The nerve endings for the retinal ganglion cells aggregate at the optical disk and form the optical nerve which transmits these signals to the brain. There are no photoreceptive cells at the optical disk, as it has a dense concentration of ganglion cell nerve endings, which are located on the inner (concave) wall of the retina. This is what causes the blind-spot in vision. Expansion and contraction of the iris/pupil pair controls the amount of light reaching the retina. Compression and dilation of the optical lens, in addition to the movements of the pupil, allow varying focus. Human vision is binocular and forward-facing, allowing a greater perception of depth.

The retina has a very pronounced curvature (see Figure 2-1) and can receive light from surprisingly high incident angles, up to 104° away from the axis of sight provided that surrounding anatomical features do not obstruct the field of view [Pirenne, 1967]. Although light at these extreme angles can be detected, it is very indistinct. This is due to photoreceptor density but also the distorting properties of the cornea, which are particularly noticeable at incident angles of 90° from the axis of gaze, perpendicular to the axis of the camera (see Figure 2-2).

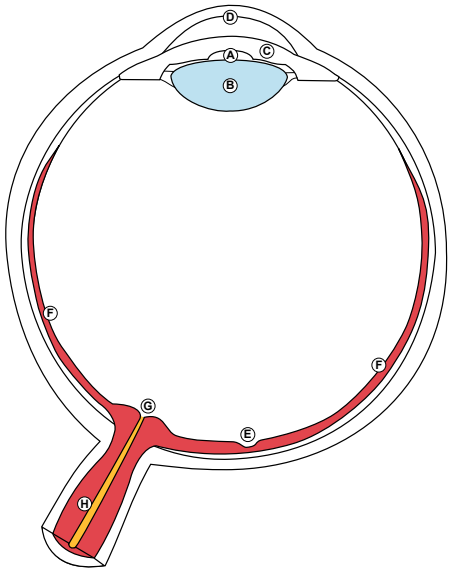

Figure 2-1: A diagrammatic cross-section of the human eye (right eye) with basic anatomical components labeled, with (a) pupil, (b) lens, (c) iris, (d) cornea, (e) fovea and foveola (central region of fovea), (f) peripheral retina, (g) optic disk (blind-spot), and (h) optical nerve. This terminology is used throughout the document.

Figure 2-2: An image of the left eye with the axis of gaze perpendicular to the axis of the camera. Corneal distortion is clearly visible, as the pupil still appears to be elliptic and facing slightly towards the camera despite gaze direction.

2.1.1 Regions of the Retina

The retina is one of the most critical structures of the eye for this research. The retina can be roughly subdivided into two regions split by the vertical meridian and two regions split by the horizontal meridian. The nasal retina corresponds to the vertical meridian half that is adjacent to the nose, while the temporal retina corresponds to the half opposite and adjacent to the temple. The superior retina corresponds to the upper horizontal meridian half and the inferior retina to the lower half. The regions are illustrated in Figure 2-3. As a consequence of this structure, the image falling on the retina is actually upside down in reference to the world and the brain corrects this further down the visual pathway.

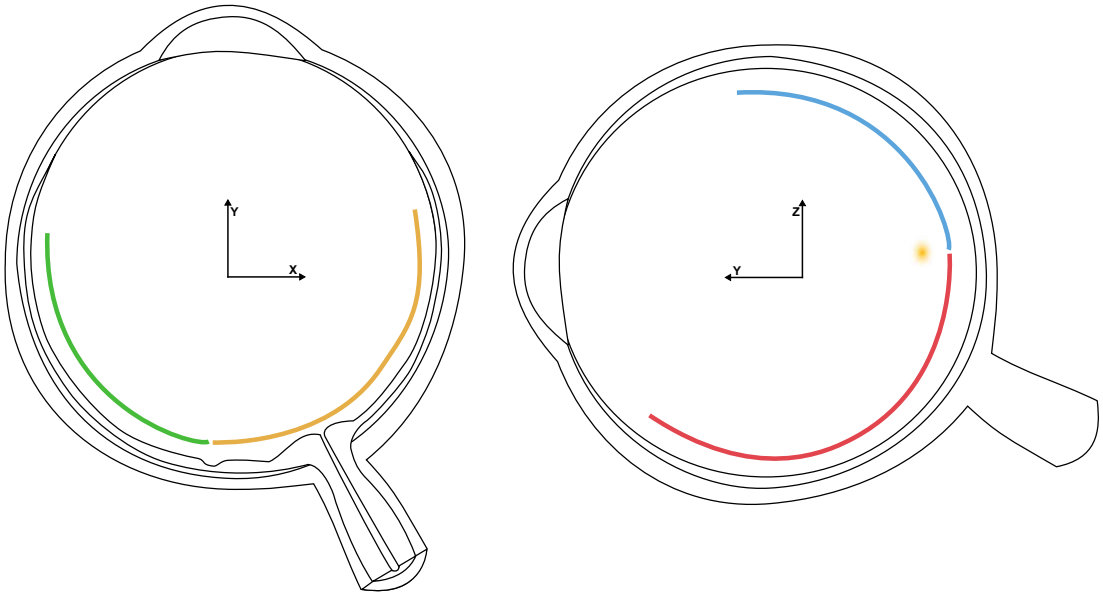

Figure 2-3: Two diagrammatic views of the (self) left eye with retinal regions indicated by color. The green region corresponds to the temporal retina, yellow to nasal retina, red to inferior retina, and blue to superior retina. Mirroring the left diagram horizontally would give the retinal locations for the right eye. The fovea is represented by a yellow gradient in the vertical cross-section for clarity.

Foveola	0.0° to 2.0°
Fovea	2.0° to 5.0°
Parafovea	5.0° to 6.7°
Perifovea	6.7° to 16.0°
Near-periphery	16.0° to 24.0°
Mid-periphery	24.0° to 40.0°
Far-periphery	40.0° and beyond

Table 2.1: Retinal regions and the areas roughly subtended on the retina. [Polyak, 1941]

The uneven distribution of photosensitive cell types in the retina contributes to differing perceptual attributes across the visual field. The fovea, and even more so the foveola, contains the highest concentration of color sensitive cells and corresponds to the region of highest visual acuity. The peripheral retina has an uneven distribution of rods and cones, but largely consists of rods. See Table 2.1 for further naming of various angular zones within the retina. These anisotropies contribute to the phenomena discussed in Section 2.2. As mentioned previously, the optic disk contains no photo-

receptive cells and thus receives no visual information. The combination of stereo vision and filtering by the brain ensure the gap is generally imperceptible.

2.1.2 Rod, Cone, and Retinal Ganglion Cells

The light sensitive biological structures of the eye are the rod and cone cells, located on the outer (convex) wall of the retina. Cone cells, responsible for colour vision, are at their highest density at the foveola and follow a pinched cone distribution, plateauing at approximately 20° of eccentricity from the foveal centre. In the human visual system, there are typically[1] three types of cone cells; sensitive to long wavelengths (red spectrum), medium wavelengths (green spectrum), and short wavelengths (blue spectrum). Wavelength (L,M,S) and colour (R,G,B) nomenclature for cone cells will be used interchangeably in this document.

Rod cells, sensitive to light intensity, are responsible for scotopic vision (vision in low light conditions). They are at their highest densities within 20° to 30° from the foveal centre, gradually decreasing with eccentricity.

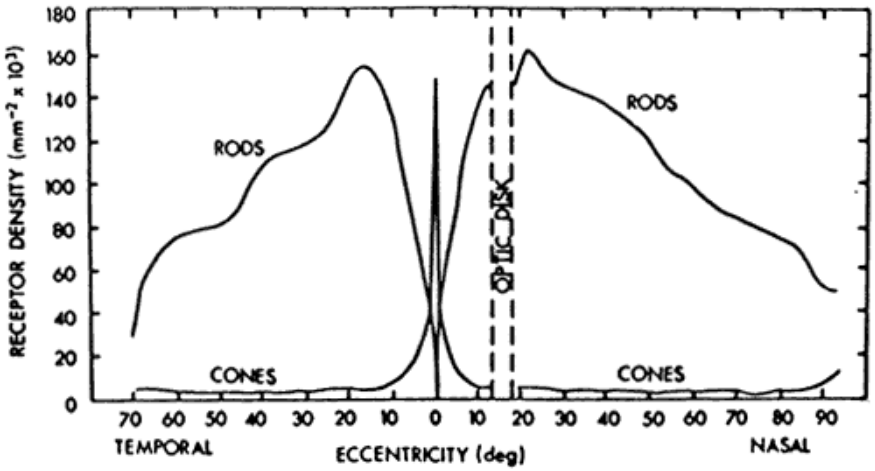

Figure 2-4: Density of cones and rods cell types across the retina. Note the sharp peak in cone cells at the fovea and the rapid drop in rod cells entering the perifovea. After [Osterberg, 1935] from [Pirenne, 1967].

[1]Tetrachromacy is a condition where an individual posses four different cone cells, and is most common among women.

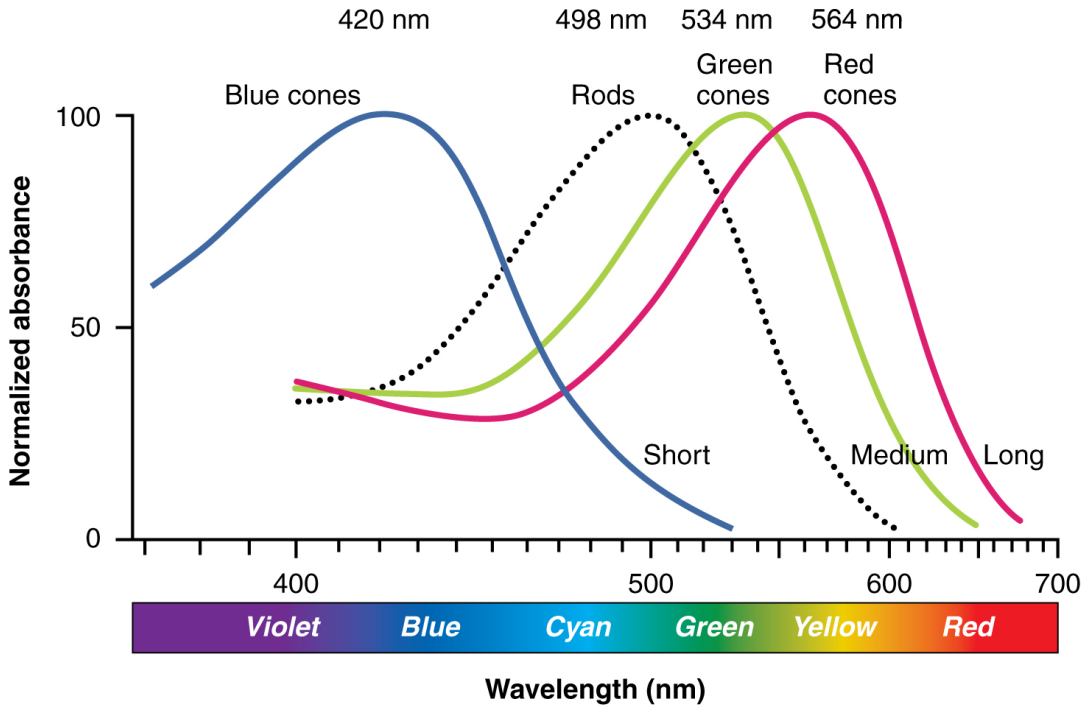

Figure 2-5: Relative spectral sensitivities of L (red), M (green), and S (blue) cone cells and rod cells under photopic conditions. From [OpenStax College, 2013] and verified against [Stockman et al., 1993]. Scotopic and mesopic lighting conditions are not included in this diagram.

Retinal ganglion cells are located near the inner (concave) wall of the retina. They are responsible for transmitting image-forming and non-image-forming information from the retina to the optical nerve and ultimately the brain. In effect, they are the last cells operating on visual information before the signals leave the eye. There are approximately 1 000 000 ganglion cells spread irregularly across the retina, with a higher concentration in the fovea where peak density can reach approximately 35 000 cells mm^{-2} [Curcio and Allen, 1990]. Since retinal ganglion cells are located between incident light and the photoreceptive cells, some minimal scattering occurs. Fortunately, the brain is good at filtering out fixed patterns in visual signals, so nerve endings and even blood vessels are not perceived. However, extra care seems to have been taken in the foveola, where retinal ganglion endings are actually pushed aside so that incident light hits cone cells directly.

The first images of a living human eye that clearly show L, M, and S cone arrangements were captured at the turn of the century [Roorda and Williams, 1999]. The relative number of L and M cones differed greatly between the two examined subjects. However, both subjects had a similarly low density of S cones, and was "not signifi-

cantly different from random in either". These are later corroborated by [Hofer et al., 2005] with a larger sample, which showed there was significant variation of L and M cones (from 1.1:1 to 16.5:1) in the majority of the subject population. Again, the distribution was seemingly random, although some subjects displayed local clumping of cones of the same type and one subject had significantly different L and M cone ratios on opposite sides of the fovea. However, S cone density and distribution was similar across all subjects.

Retinal topography results conducted on seven deceased human retinas (six donors, all female) [Curcio et al., 1991] provide further detail on retinal structure. Approximately 7 % of cones within a 4mm radius of the foveal centre are S cones. The density of S cones increased with eccentricity up to 1° out from the foveal centre, yet they were also completely absent in a zone up to approximately 0.35° out from the foveal centre.

A study on the detectability of deep-blue or deep-red light confirms that it is a function of its eccentricity from the foveal centre, or the stimulus' distance from the point fixation [Pirenne, 1967]. With a deep-blue stimulus, there was a noticeable drop in detectability beyond 0.75° eccentricity with the subject reporting that it the stimulus appeared colourless. Conversely, the detectability of a red stimulus remained somewhat constant throughout the same distances, with a noticeable drop beyond approximately 8° of eccentricity.

Cone type densities have also been analysed across the whole retina (rather than just the fovea) from a pool of 23 deceased donors, varying from 5 to 96 years of age [Hagstrom et al., 1998]. The study reported that, on average, the central retina was approximately comprised of 40 % M cones, down to around 25 % in the far peripheral retina. In prior experiments observing color stimuli, subjects had perceived green-yellow colours as orange in those regions of the retina, explained by the lowered sensitivity to green wavelengths. They observed a relatively constant ratio of L and M cone densities from the foveal centre to approximately 20° of eccentricity.

2.1.3 Saccades and Smooth Pursuit

There are two types of occulomotor movement conducted by the eye; saccades and smooth pursuit. Smooth pursuit, as the name implies, allows tracking of visual stimulus through a smooth shift of gaze. In contrast, saccades are bullet-like movements that shift fixation quickly. Interestingly, smooth pursuit requires a tracked target to function and degenerates to saccadic steps in the absence of one[2].

[2]You can verify this yourself with a simple test. Hold your finger against the edge of the table, fixating on the tip of the finger, and move it slowly in one direction along the edge. Your eyes will smoothly track your finger, occasionally saccading to catch-up if you're moving too quickly. Now

Saccadic velocity is typically proportional to the amplitude of movement. Maximum velocities can reach speeds of approximately $600\,^\circ\,\mathrm{s}^{-1}$ for amplitudes up to around 40° or even approximately $800\,^\circ\,\mathrm{s}^{-1}$ for amplitudes close to 75° [Baloh et al., 1975, Bahill and Stark, 1979, Henriksson et al., 1980]. Typically at such high speeds, the eye tends to overshoot its target and then performs subsequent corrective saccades or glissades (slow and smoother movement) to the correct fixation. There is some lack of consensus on weather temporal or nasal saccadic movement is faster, but differences are generally only significant at larger amplitudes [Baloh et al., 1975, Abrams et al., 1989, Boghen et al., 1974]. However, most naturally ocurring saccades have amplitudes within 15° [Bahill et al., 1975], which roughly corresponds to typical maximum velocities of $200\,^\circ\,\mathrm{s}^{-1}$ to $300\,^\circ\,\mathrm{s}^{-1}$.

2.1.4 Accommodation

Accommodation is the process by which the eye adjusts or maintains focus. It is comprised of three coordinated operations: the vergence of the eyes, contraction/relaxation of the ciliary muscles and thus alteration of the ocular lens shape, and a dilation/constriction of the pupil. For the purpose of this work, we will be able to detect vergence and pupillary dilation/constriction, from which we may infer the focus of the eyes and provide more accurate depth-of-field effects. Traditional methods that use eye tracking data simulate depth-of-field effects by relying solely on gaze point relations to objects within the scene. This method, however, would be unable to accurately simulate depth-of-field for multiple objects at similar positions but different depths or situations where focus is intentionally altered from what would be expected given the scene content. Changes in accommodation are relatively slow compared to other operations of the eye, taking approximately 350 ms on average to change from near-to-far or far-to-near focus [Campbell and Westheimer, 1960].

2.1.5 Cortical Magnification Factor

As we approach retinal eccentricities, not only do cone cells become sparser but the amount of visual cortex dedicated to each degree of visual field decreases. Thus, near 0° of eccentricity from the fovea there is four times more visual cortex dedicated to an area subtending a single degree of the retina than at 15° of eccentricity, and eight more times than at 25° of eccentricity [Cowey and Rolls, 1974]. This is theCortical Magnification Factor (CMF), typically represented by M, which underlies many of the perceptual phenomena described in Section 2.2. Although it may appear from Cowey

remove your finger and try to do the same thing by just looking along the edge of the table. Your eyes will saccade continuously, with no smooth pursuit occurring.

and Rolls' results that there is no anisotropy of M across the retina, direct cortical measurements of M provide values for binocular vision. Rovamo and Virsu [Rovamo and Virsu, 1979] elucidate that for monocular vision there are still asymmetries of M across the retina.

2.2 Visual Psychophysics

Visual perception is achieved through a combination of data received and encoded by various structures of the eye and its subsequent representation in the brain. Through the process of natural selection, many parts of the perceptual system had to compromise on particular aspects. For example, if the entirety of the human retina were to have the same density of cone cells as the central fovea, the optic nerve and brain would have to engorge to absurd amounts to accommodate the influx of information. To compensate, our visual system has evolved to preserve high acuity (spatial) only at the centre of gaze while maintaining good perception of motion (temporal) within the periphery.

An in-depth understanding of existing phenomena allows the derivation of better computational models for rendering. However, an complete understanding of all existing visual phenomena would be of very limited use for this work. There are many phenomena that are have only a minimal effect (for the purpose of this research) on perception, and therefore it is hard to see any use in their exploitation in practice. Conversely, there are also some that require more extensive study before we can determine their utility, and occasionally their effects may be conflated with other, similarly misunderstood, phenomena. The following is not an exhaustive list of all visual phenomena, but a selection determined to have enough relevance and supporting literature to be worth exploring. Each section will provide a brief description of the phenomena and some key observations that will help develop our models.

2.2.1 Flicker Fusion

Flicker fusion is the phenomenon in which, at a certain critical frequency, distinct and sequential pulses of light appear as one continuous light source. It is a well established phenomenon that is widely exploited, from room illumination to electronic displays, and are generally designated as strobe lighting. In contrast, constant lighting solutions remain lit continuously. In order to match the levels of apparent brightness of a constant source, strobe sources flash brighter in order to compensate for lack of lighting between each strobe, thus normalizing the apparent brightness. By maintaining the strobing frequency above the Critical Flicker Frequency (CFF) we can emulate a continuous light source. It is important to note that this sections solely addresses

the detectability of light source flicker, and not the illusion of apparent motion, which is governed by a separate phenomenon.

The phenomena has been studied for over a century, and for most of this time studies appeared to produce incongruous results (see [Lythgoe and Tansley, 1929, Hecht and Verrijp, 1933] for an early example). The apparently conflicting results were mainly due to inconsistent experimental conditions and sometimes a lack of rigorous control for known variables. Since the body of literature on the subject is long, extensive, and at times confusing, I have condensed the material into a few general insights with little in-depth discussion. For an excellent summary of flicker fusion research, along with additional references, see Section 3.6.3 in [Strasburger et al., 2011].

Any discussion of Critical Flicker Frequency (CFF) will bubble up two key observations. Firstly, CFF thresholds depend on the luminance of the stimulus signal. This is known as the Ferry-Porter law (Equation 2.1) [Ferry, 1892, Porter, 1902].

$$f = \alpha \log L + \beta \tag{2.1}$$

Where f is the CFF, L is the stimulus luminance, and α and β are constants. Secondly, CFF thresholds depend on the size of the stimulus signal. This is known as the Granit-Harper law (Equation 2.2) [Granit and Harper, 1930].

$$f = \alpha \log A + \beta \tag{2.2}$$

Where f is the CFF, A is the area subtended by the stimulus in degrees, and α and β are constants. These observations were originally evaluated against stimuli that appeared at or near the point of fixation (within the fovea). However, several later studies have shown that the eccentricity of the stimulus from the point of fixation (subsequently its location on the retina) also affects the CFF threshold. This had already been suggested by Granit and Harper in their study, but was only solidified as future studies explored stimuli on regions of the retina beyond the foveal boundary. After an exhausting number of years and conflicting results, it seems that it could clearly be said that the CFF threshold is affected not only by luminance and stimulus size, but also the eccentricity of the stimulus in relation to the luminance and size. To be succinct, the following critical observations can be made:

Observation 1 – As any single light stimulus increases in luminance but remains constant at a given size and eccentricity, CFF tends to increase (revised Ferry-Porter law).

Observation 2 – As any single light stimulus increases in size but remains constant at

a given luminance and eccentricity, CFF tends to increase (revised Granit-Harper law).

Observation 3 – As any single light stimulus varies in eccentricity...

(a) but remains constant at a given luminance, **CFF tends to decrease with eccentricity if the stimulus size is small** (<3°), plateauing at the mid-periphery [Alpern and Spencer, 1953, Ross, 1936, Creed and Ruch, 1932, Fukuda, 1979].

(b) but remains constant at the same given luminance, **CFF tends to increase with eccentricity if the stimulus size is large** (>3°), plateauing at the mid-periphery [Riddell, 1936, Hylkema, 1942, Tyler and Hamer, 1990, Fukuda, 1979].

Observation 4 – For any single light stimulus of any given luminance, size, and eccentricity, CFF is raised if the stimulus is surrounded by a brighter background [Creed and Ruch, 1932, Fukuda, 1979].

Observation 5 – For any single light stimulus of any given luminance and reasonable size, CFF in the periphery tends to decrease as eccentricity approaches the far-periphery.

A revision of the Ferry-Porter and Granit-Harper laws that takes eccentricity into account was later later formalized by Poggle et al (Equation 2.3) [Poggel et al., 2006].

$$f = (\delta E + \alpha)(\log L + \gamma \log A_E - \sigma E) + \beta \tag{2.3}$$

Where f is the CFF, E is the eccentricity of the stimulus in degrees, L is the stimulus luminance, A_E is the area subtended by the stimulus at eccentricity E in degrees, and δ, α, γ, σ, and β are constants. By accepting these observations, we can account for the common experience that 50Hz strobe light sources (like CRT monitors) appear to flicker in peripheral vision without discounting a large portion of existing literature.

2.2.2 Separable Acuity and Contrast Sensitivity

There are several methods of measuring visual acuity at the fovea. Snellen charts, which are commonly seen at optometrists, are one example, although Landolt-C tests tend to be better than Snellen letter identification. We are more interested in generalizable measures of acuity, particularly those relevant to the hardware we use for current

and future work. Much like camera sensors have a limited number of individual photon collecting cells the human eye also has a limited number similarly small photo-receptive cells, and details that subtend areas smaller than the cell size are inseparable. This is Minimum Separable Acuity (MSA), the smallest distance at which two stimuli can be discriminated from one another [Sanders and McCormick, 1987]. Of course, the size of individual retinal cells and their spacing are not as easily determined as citing a camera sensor's manufacturer sheet.

Minimum Separable Acuity (MSA) highly depends on stimulus contrast between and the luminance level of the targets and the background. Under laboratory conditions with close to 100 % contrast (black and white) and approximately $350\,\mathrm{cd\,m^{-2}}$ subjects were able to discriminate the stimuli down to a distance of $0.5'$. At 50 % contrast and approximately $35\,\mathrm{cd\,m^{-2}}$ the smallest separation that could be discriminated was $1.0'$ [Curry et al., 2003].

MSA does not represent the highest level of performance possible for the human eye. Hyper-acuity measures refer specifically to the detection of the position or presence of an object, however these acuity measure are of little relevance to this research. MSA allows us to calculate the minimum pixel size/spacing required such that any further refinement would yield no perceptual benefit. That is, the pixel density (center-to-center spacing) required at a particular viewing distance such that pixellation is undetectable (Equation 2.4).

$$D_{pixel} = D_{display} * 2\tan(\frac{\alpha}{2}) \tag{2.4}$$

Where p is the individual pixel size, d_{disp} is the distance between the eye and the display, and α level of acuity desired in radians. We have mentioned that MSA is sensitive to both luminance and contrast, so we should assume acuity values that take the display's maximal contrast and overall luminance into account. Due to this dependence on contrast, acuity is also typically measured by determining the minimal amount of contrast needed to discern two stimuli apart over a range of spatial resolutions, known as a Contrast Sensitivity Function (CSF).

Contrast sensitivity is typically measured using achromatic sine-wave gratings and varying the cycles per degree (spatial frequency) and contrast between the peaks and troughs of the function (black to white at highest contrast). Contrast sensitivity versus spatial frequency plots tends to follow a dipper shape [Legge and Foley, 1980, Legge and Kersten, 1987, Stromeyer and Klein, 1974, Wilson, 1980]. As frequency increases (more cycles per degree), contrast sensitivity increases slowly at first and then rapidly begins to decrease.

Studies have shown that there is no significant difference in the shape of the Contrast

Sensitivity Function (CSF) at different locations in the retina [Virsu and Rovamo, 1979] but there is a difference in scale. When the stimulus size is constant and not scaled to account for cortical magnification, sensitivity scales with retinal eccentricity (as eccentricity increases, so does the threshold), but not the overall shape of the function [Legge and Kersten, 1987]. However, there is a nasal-temporal retinal asymmetry with higher performance in the nasal retina [Anderson et al., 1991]. This performance difference is most notable at and beyond 20° of eccentricity. Performance for chromatic stimuli was lower, but followed the same pattern of retinal asymmetry. There was no apparent asymmetry at the vertical meridian.

The contrast sensitivity and acuity measurements so far all refer to photopic vision. Performance for mesopic and scotopic vision is much poorer, but follows similar patterns (with the exception of any significant retinal activity in the fovea). We will not be dealing with scotopic lighting conditions throughout this research and, given that photopic vision is of much higher quality, any perceptually-lossless rendering adjustments made would be inherently refined enough to account for mesopic/scotopic vision. To summarize:

Observation 6 – Contrast sensitivity is at its highest at an intermediate value of spatial resolution.

Observation 7 – Contrast sensitivity in the temporal retina is significantly poorer than in the nasal retina once the eccentricity of the stimulus exceeds 20° [Anderson et al., 1991].

Observation 8 – Contrast sensitivity for chromatic stimuli is significantly poorer than for achromatic stimuli.

2.2.3 Saliency

The human visual system is also naturally attracted to specific kinds of stimuli. Studies have found that human gaze is aimed, even during casual observation, towards areas with irregular contours or unique areas [Mackworth and Morandi, 1967]. For a subjective level of informativeness, subjects had a significantly larger number of fixations on the highly informative features of an image in comparison to regions with low or average detail. Mackworth and Morandi suggest that uninteresting and predictable features of the image are processed peripherally and screened such that the fovea is only directed to novel or unpredictable stimuli. To support this, they clarify that most of the unusual and informative areas of the image in their study were fixated within two seconds of the start of the presentation, suggesting that peripheral vision assisted by editing out the uninformative regions.

Visual attention and saliency is often decomposed as the result of two visual mechanisms operating in unison. **Top-down saliency**, or goal-driven, refers specifically the saliency of image features based on their relevance to the task at hand or the subject's own internal intentions. **Bottom-up saliency**, or stimulus driven, refers to the natural saliency of certain image features, such as regions with harsh edges or contrast, or objects in motion. This is the form of saliency addressed in Mackworth and Morandi's work mentioned above. Although both forms merit study, it is generally easier to investigate and conceptually easier to exploit bottom-up saliency.

For example, it has been found in a free-viewing experiments that fixations tend to lie on high contrast patches of the image or in regions with object edges [Reinagel and Zador, 1999]. With prolonged observation, consistency of fixed locations between different observers decreased [Tatler et al., 2005]. Prior studies confirm these tendencies, suggesting that during the first second and a half of presentation, correlation between gaze fixations of several different participants on the same image was very high, weakening as the exposure time increased [Mannan et al., 1995].

This suggests that viewing novel imagery may be dominated by bottom-up mechanisms for the first moments of exposure, and top-down mechanisms beyond that. However, top-down mechanisms can exert a lot of control over visual attention, especially in imagery that excite cognitive surprise [Loftus and Mackworth, 1978]. Ultimately, saliency can be though of a competition between both mechanisms for control of visual attention. To summarize:

Observation 9 − Visual attention and gaze is governed by both top-down (goal-driven) and bottom-up (stimulus-driven) features of a stimulus.

Observation 10 − When presented with a novel stimulus visual attention tends to be dominated by bottom-up features, such as areas with large variability in contrast or edges, for the first few moments of observation.

Observation 11 − Visual attention tends to be dominated by top-down features, such as task-oriented features, cognitive surprises, or internal preferences, after the first few moments of observation.

Observation 12 − Visual attention on bottom-up features tends to be consistent across all participants for any given image.

2.2.4 Directing Gaze

So far we have only reviewed passive phenomena. Directing the user's gaze through subtle or unnoticeable distractions is an active alternative. Some recent work has

studied the effect of image modulation to direct user's gaze in a simple counting exercise, evaluating task performance with subtle modulations, overt modulations, and a control with no modulation [McNamara et al., 2008]. These modulations were only applied in the periphery, and were terminated as soon as the user's gaze began moving towards the modulation. Subtle modulations were only $2.4'$ in size, while overt modulations subtended $7.5'$. Performance for both modulated trials exceeded that of the control. All users that conducted the overt modulation trials found the modulations distracting but useful, helping them find objects they wouldn't otherwise have noticed. However, participants in the subtle modulation trials did not notice any of the modulations and their average performance was as good as the overt modulation trials, suggesting that the only difference between the modulated trials was the level of disruption.

Other investigations have found similar results; a subsequent study found that their users, although also unaware of the modulations, perceived static scenes to be of higher quality than their modulated counterparts [Bailey et al., 2009]. However, modulated areas in their trails were manually selected to appear in locations that were not visually significant. These locations had low contrast, low detail, low colour saturation, or generally contained uninteresting objects. It is possible that the content users are directed to, rather than the modulations themselves, are the cause of the perceived loss of overall quality.

Additionally, modulations required an activation time before movement in the direction of the modulation began. Activation time was within 0.5 seconds for approximately 75 % of the participants, and within 1 second for 90 %. Additionally, 69 % of gaze directed movement endpoints fell within 160 pixels of the modulated target, with 93 % falling within 320 pixels.

Directing the user's gaze provides us with an active route for perceptual manipulation. As we have seen, additional considerations may be required when using such a method to avoid redirection to poorly chosen areas. On the other hand, this suggests that we may able to enhance the perceived quality of an image simply by redirecting gaze to highly informative or interesting points in the scene (based on the metrics discussed in the saliency section). To summarize:

Observation 13 − It is possible to direct the user's gaze through undetectable modulations, provided the modulations are spatially small and are immediately terminated as the user's gaze begins to move in the direction of the modulation.

Observation 14 − The overall perceived quality of an image may be affected by the locations the user's gaze is redirected to.

2.3 Eye Tracking

High quality eye tracking is possibly the most critical and sensitive portion of this work. Failing here means that any work evaluating our novel perceptual models may be dismissed due to failures in the tracking system. We are not interested in developing a novel tracking algorithm with this work; instead we hope to integrate as much as possible from existing systems and construct our own system based on this prior work.

Our primary concerns are speed, accuracy, precision, and informativeness of the tracking method. It is preferable if the system can easily be adapted for virtual reality environments and/or head-mounted display. We are forgoing typical desk-mounted or remote tracking solutions for these reasons. Instead, we have opted for a head-mounted solution with cameras directed at both eyes in order to achieve the highest resolution quality possible. This section only addresses the most relevant contributions in the field to this work, specifically. Consider it a selective depth-first exploration of the literature.

We are interested in capturing two key features of the eye; firstly, the centre of gaze to track fixation and secondly, the pupil's dilation to estimate focus. We are unaware of any method that would allow capture of lenticular dilation/contraction using only the hardware that is commercially accessible to us, so we aim to use pupil dilation as an estimate for focus. Both of these features are provided by accurate segmentation of the pupil.

Capture under visible light presents a few difficulties for pupil segmentation. For example, the cornea will reflect incoming point light sources and occasionally more defined imagery. Dark irises can also interfere with the detection of pupillary boundaries. We also might want better control of the lighting conditions in the eye socket, and visible light is often a significant disturbance to the user.

In contrast, Near-Infrared (Illumination) (NIR), typically defined within the 700 nm to 1400 nm spectral range, overcomes these problems. Capturing in Near-Infrared (Illumination) (NIR) means that most uncontrolled light sources that are reflected by the cornea are blocked and no longer present in the imagery. Dark irises also appear lighter and more detailed, allowing a better differentiation at the pupillary edge. Most importantly, depending on the angle of NIR source, the pupil appears as pitch black [Myers et al., 1991] or bright white [Hutchinson et al., 1989], and as NIR is outside of the visible spectrum we have significant control over illumination without disturbing the user.

There are many methods which take advantage of infra-red illumination to track the dark/bright pupil, or at times both [Ebisawa, 1998]. In 2005 the Starburst algorithm

was introduced [Li et al., 2005] that used dark pupil NIR illumination and an ellipse fitting algorithm, similar to prior work [Ohno et al., 2002]. The Starburst algorithm operates by casting a given number of rays radially outward from an estimate centre. Travel along each ray is halted when a candidate pupil edge point is reached, and a feature point is defined. The algorithm then repeats the ray casting for each feature point but from a restricted casting angle, such that if the feature point lies on a pupil edge casting rays in that restricted angle will generate feature points on the opposite side of the pupil. A new estimate centre is calculated and the processes is repeated until convergence or a limit of iterations.

This feature point set contains a lot of noise. Many points will be on eyelid edges, eyelashes, and other naturally dark features. The algorithm then refines the feature point set using RANSAC [Fischler and Bolles, 1981] to carry out ellipse fitting. Once an ellipse is found that fits against a majority of the feature points, or a limit of iterations is reached, the ellipse is returned.

Later adaptations and extensions suggest relatively poor performance of the algorithm in non-ideal conditions (e.g. off-axis imagery). A later paper detects elliptical fit pairs on the pupillary and iris boundaries, reporting a significant improvement over the Starburst algorithm [Ryan et al., 2008]. Świrski et al. present a novel algorithm that shares some similarities to Starburst [Świrski et al., 2012]. Their method uses Haar-like features to extract a rough region corresponding to the pupil, followed by a morphological open and Canny edge detection. The RANSAC-based ellipse fit is then carried out on the feature points detected by Canny edge. Their results suggest a significantly higher detection rate over Starburst and good performance in highly off-axis images and images with a lot of eyelash interference.

2.4 Displays and Virtual Reality

In 1965, Ivan Sutherland proposed that computers of the future would be capable of sensing the position and motion of muscles, including motions of the eye to display information dependent on gaze, and eventually control the existence of matter within a confined space [Sutherland, 1965]. We are far from the ultimate computer assisted room that Sutherland imagined, but the inklings of human aware computing are here. Out of all the fields within computer graphics, those specific to virtual reality are perhaps the (or at least, should be) most conscious of passive human data collection and the most perceptually aware. Some of the earliest examples of immersive technologies were already deeply aware of the importance of providing multi-sensory feedback [and others, 1962].

Immersive technologies can be binned into intrusive and non-intrusive categories. Intrusive technologies, such as head-mounted displays or haptic jackets, require direct attachment to the user and generally serve as sensory replacements. Non-intrusive technologies, such as computer assisted virtual environments or domed projections, rely on adjusting the user's surroundings and indirectly fooling the senses. The following sections will briefly go over historically significant work in immersive technology and then focus on state of the art that explicitly track's user data and exploits perception. Concerns and shortcomings of existing technology will be made clear to the reader, lending to the solutions proposed in this thesis.

2.4.1 Head-Mounted Displays

Recently there has been a revival of commercial interest in Head-Mounted Display (HMD) technology. Largely instigated by the popularity of the Oculus Rift, there is a push towards bringing head-mounted display technology out of the academic and military environment and into the hands of consumers. This has led to the development of even more specialized and competing devices with wide field of view, integrated eye tracking, or novel position tracking systems[3]. Despite individual shortcomings, the variety of choice in the HMD shows there is a vibrant and growing community in the field of personal VR with many issues that are still to be addressed.

2.4.2 Computer Assisted Virtual Environments

In contrast to HMDs, a Cave Automatic/Computer Assisted Virtual Environment (CAVE) relies on altering what is displayed in the user's environment, usually on enclosing walls or encompassing surfaces, to yield the illusion of presence. Cruz-Neira et al. [Cruz-Neira et al., 1992] developed one of the first Cave Automatic/Computer Assisted Virtual Environments (CAVEs) to be discussed in an academic setting. Their report brings up some of the advantages CAVE-like systems have over other VR systems, such as full-body immersion and, theoretically, multi-user support in the same physical space. Although their initial work was unable to account for moving users, their system laid the groundwork for collaborative virtual environments of the future with little to no intrusion on the user. More recently, the CAVE2 [Febretti et al., 2013] uses 72 LCD panels in a cylindrical arrangement to provide a large space Virtual Reality Environment (VRE). Since then, there have been many other implementations of CAVE-like systems using display panels and projectors. However, this thesis will not

[3]For a few examples, see StarVR (http://www.starvr.com), FOVE (http://www.getfove.com), and HTC Vive (http://www.htcvr.com).

be going into historical details and will instead focus on some of the problems innate to typical display systems.

2.4.3 Pixel Persistence

Displays can also be categorized based on the time individual pixels remain lit. In sample-and-hold displays, a pixel persists for the entire duration of the frame cycle time and pixel switch delay and therefore the light emitted from the display is continuous. Alternatively, displays with a low persistence only remain lit for a portion of the cycle and switch time, producing a strobing emission. Low persistence displays require higher brightness emissions for the duration they are lit in order to be perceivably the same brightness as an equivalent sample-and-hold display.

At currently average 60 Hz framerates, there is practically no difference between either technology when displays a static scene. However, a stark difference appears as the eye begins tracking dynamic objects in a scene. At full persistence, the object will appear to blur as the eye continues to move in the direction the object is supposed to be in and integrates the static image being continuously displayed for the duration of the frame. The solution to this is lower persistence, as the image is only integrated for a shorter duration of time, thus reducing blur. However, low persistence can cause a strobing effect in which the object appears to jump from one location to another without moving in between (which, in essence, is what it is doing). The solution to both of these problems is a higher content refresh rate, as the eye integrates less of a single static frame and the inter-frame jumps become imperceptibly smaller. Of course, this still depends on the velocity of the tracked object.

An in-depth study by Kuroki et al. [Kuroki et al., 2007] found that content refresh rates of 250Hz were high enough to eliminate blur and jerkiness/strobing with a 480Hz display panel. Although they used a CRT monitor (which is inherently not full-persistence technology), their experiments simulated sample-and-hold displays by repeated display of the same frame during the lower FPS simulations (so although the emission rate remained constant at 480Hz, the content rate was lowered to the simulated rate). Additionally, the used frame averaging to reduce their 1000 FPS source footage to the rates they were simulating. Unfortunately, this casts some doubt on their low refresh rate results for high speed objects, as the perceived blur may have been affected by the footage's blur (in contrast to rendered imagery which remains sharp between each frame when no post-processing or temporal anti-aliasing blur is applied), but for higher frame rates their results will have been more accurate. Therefore, it is reasonable to assume 250Hz is a good starting point to aim for in order to resolve strobing and blur for typical object speeds.

2.4.4 Bezels

Bezels are an inherent problem with any non-projection based display technology. Although seamless or close to seamless display configurations are available, these systems typically cede on image quality at the seams or other properties of the display to lower cost. As such, bezel compensation methods are still relevant to study for multi-monitor configurations. There are several approaches that can be employed in such systems.

Although there is an obvious aesthetic issue that remains unresolved, several studies have shown that task performance is not significantly different between (and can sometimes benefit from) tiled versus continuous multi-display setups. McNamara et al. [McNamara et al., 2011] show no significant difference in performance during a navigation task between tiled and continuous displays. Grüninger and Krüger [Grüninger and Krüger, 2013] surveyed the effect of bezels on stereoscopic vision and found that bezels, although significantly degrading overall depth perception, did not interfere with counting tasks. Tan and Czerwinski [Tan and Czerwinski, 2003] found that "physical discontinuities introduced by bezels as well as by differences in depth alone do not seem to have an effect on performance", although the tasks their users performed were located entirely within the visible areas of the canvas. Hennecke et al. [Hennecke et al., 2012] also report that there is no significant difference between curved, edged, or bezeled screen transitions (although trials on curved transitions were more accurate).

However, for our purposes aesthetic degradation is a critical issue and must be compensated for. There are four general methods to deal with bezels. Firstly, the offset model, assumes display regions are connected as if the bezel were not present. The advantage of this method is that no information is lost, but there are visible discontinuities in the image. Secondly, the overlay model, assumes the bezels overlay a larger, underlying image that spans the area corresponding to the entirety of the display surface and bezel widths. In this case, discontinuity is partially resolved but information is lost. These first two methods do not deal with bezels in a satisfactory method, and are suited for scenarios where bezel can be ignored.

Thirdly, the french-window model, assumes the image lies on a different layer than the display surface. This method requires head-tracking or some form of control so that the user can look around or behind bezels, thus avoiding discontinuity and information loss at the cost of system complexity. Almeida et al. [De Almeida et al., 2012] present two solutions, named ePan and GridScape, that adopt this model. For ePan, users are able to drag the underlying image with gestures. More interestingly, GridScape displaces the canvas further "behind" the display and tracks the user's position to produce a motion parallax effect [Gibson et al., 1959]. They evaluated both methods

against the base offset model, which inherently does not suffer from obfuscation, in a simple path tracing trial. Almeida et al. found that, although there were more errors with the offset method, the difference in errors between the methods was not significant. Subjectively, users found GridScape to be "more intuitive" and "more fluid" than the other methods. In situations where eye or head tracking is already present for other reasons, such as foveation, it seems that a model similar to GridScape for bezel compensation (if necessary) would be subjectively better.

Lastly, the projection model, uses external hardware to project imagery that would be hidden by the overlay method onto the bezel surfaces. The model is highly sensitive to alignment on all three dimensions (bezels tend to be elevated further than display surfaces and must be compensated for accordingly), but does not suffer from obfuscation or discontinuity. Ebert et al. [Ebert et al., 2010] found that users preferred this method over the offset and overlay models.

2.4.5 Registration

Seamless imagery with multiple projectors is a different problem. In this case, there is no longer a problem concerning bezels, but rather proper alignment and calibration of the projectors. Registration is the process in which image features on a target display surface across multiple projectors are aligned (geometric registration) and blended (photometric registration) correctly. This is a mixture of the proper configuration of the projector settings and imagery, the geometry of the display surface, and (when stereoscopic effects are concerned or the user can obfuscate the projections) the position of the user.

Typically, multi-projector based systems assume the display surfaces are fixed and calibration occurs *a priori*. Additionally, the user's position is assumed to be fixed for stereoscopic effects. This is the case of the original cAVE [Cruz-Neira et al., 1993], where non-overlapping rear-projection on the surfaces also ensured there would be no obfuscation. For this body of work, we are not interested in scenarios where the display surface is dynamic and so we will circumvent that literature. We are, however, very interested in ensuring dynamic blending and adjusting for stereoscopics.

Raskar et al. [Raskar et al., 1999] supplied convincing results for their geometric and photometric registration methods. Their system registers three projectors successfully and uses two cameras for the *a prior* calibration process for surface estimation. Later work by Raskar et al. [Raskar et al., 2006] applied registration to quadratic surfaces such as domes. The RoomAlive system [Jones et al., 2014] is a recent example of registration along complex surfaces. Jones et al. use Kinect depth sensors to calculate the room's geometry.

Nakamura and Hiraike [Nakamura and Hiraike, 2002] introduced a geometric registration method for dynamic pan-tilt projection using a single camera for calibration. The surface geometry is captured *a priori* with a rigidly attached camera that captures several stills of a projected calibration pattern. Ashdown and Sato [Ashdown and Sato, 2005] also supplied their own calibration method with more in-depth detail. Their system uses a pan-tilt mirror to redirect the projection, rather than actuating the projector itself. Their method also uses a projector camera pair, and calibrates using an iterative pose estimation process with a projected calibration pattern. The system only seems capable to handle flat surfaces as there is no mention of complex geometric registration. Dynamic systems like these have to account for both the distortion caused by the surface geometry and the angle of projection (keystone effect).

2.4.6 Pan-Tilt Projection Steering

Of course we can rely on sufficiently high update cycles for projectors and use them much in the same way we would other display hardware. But, unlike 120 Hz to 240 Hz monitors which are now more commonplace and commercially accessible, projectors that meet the same specifications while also providing high resolution and low display latency are inordinately expensive. However monitor based systems suffer from a significant setback: structural rigidity. It is difficult to change the physical configuration of a monitor array without significant physical engineering effort, but with projectors it boils down to clever initial arrangement of the units and the use of mirrors to redirect the projections. Registration and proper render projections for the new configuration and surface are then handled by software, using supporting hardware (e.g. a witness camera) when needed.

An additional advantage of a pan-tilt projection systems is that there is an immediately evident advantage to their use for gaze-contingent rendering. By combining two consumer level projectors and one pan-tilt mirror system, one static projector can be used for the peripheral render while the second dynamic projection (using the mirror system) is redirected to the user's gaze. In this way we also simulate the quality of the foveal projector over the amount of surface coverage provided by the peripheral projector. This method introduces some additional benefits due to direct control of the foveal region's hardware-bound aspects (e.g. projection area and, consequently, pixel density) but at the cost of overall increased system complexity, since the method requires active distortion correction and rapid registration to account for redirection.

An early example of a system that redirected a projector on pan-tilt and corrected for redirection and surface distortions is that of Nakamura and Hiraike [Nakamura and Hiraike, 2002]. Although relatively fast, as the system had to rotate an entire camera-

projector paired system, it could not achieve the speeds required for saccadic movement. Mitsugami et al. describe a calibration process for a steerable projector system with a fixed projection/rotation center [Mitsugami et al., 2005]. Staadt et al. introduced a foveal pan-tilt steerable projection system for high resolution data observation [Staadt et al., 2006], but their system is restricted to flat projection surfaces and is not capable of saccadic speeds.

In 2011, Okumura et al. developed a pan-tilt mirror system that could achieve saccadic speeds [Okumura et al., 2011]. Although the initial purpose of their system was for high speed tracking, they later extended the work with an active projection mapping component that simultaneously tracked and projected onto a target object [Okumura et al., 2013]. Redirecting the projection itself requires the use of additional optics as well as more distortion correction, but allows for saccadic active projection due to the light weight of the actuated mirrors. Ashdown and Sato describe a calibration process for a steerable projection system [Ashdown and Sato, 2005].

In using two projectors we also need to provide accurate active projection blending and alignment. Chen et al. introduced a method to automatically align projection surfaces for static layouts [Chen et al., 2000]. In 2001, Yang et al. introduced a reconfigurable multi-projector display system called PixelFlex [Yang et al., 2001]. Using eight projectors, their system could change to an arbitrary static display layout and would automatically calibrate the new configuration. Brown et al. provide a survey of many geometry and color registration methods, including ones for non-planar surfaces [Brown et al., 2005]. In 2009, Sajadi et al. introduced a novel automatic calibration method with perception based constraints that accounted for variation in chromaticity, vignetting, and overlap blending [Sajadi et al., 2009].

2.5 Rasterization and Ray-Based Rendering

Real-time graphics has been dominated by rasterization methods for many years. The advent of dedicated graphics hardware (particularly NVIDIA's GeForce 256, marketed as the first Graphics Processing Unit (GPU) with integrated hardware engines for transform, lighting, triangle set-up and clipping, and rendering operations) further solidified their dominance within interactive applications. These models make several approximations and "good enough" hacks that allow them to be much faster (but also more physically inaccurate) than ray-based methods.

These approximations are its biggest drawback, however, as they do not provide accurate representations of how light operates in the real world. A simple example of this are shadows; rasterization methods require abstract techniques like shadow maps

where shadows are inherent to ray-based methods. Compensating for these abstractions often requires costly computational workarounds that push against interactive framerates. Although many advances have been made in the past decades, state-of-the-art real-time rendering still leaves much to be desired when compared to respectively modern ray-based methods.

As such, there are many reasons why we would prefer to (but are not limited to) operate with ray-based methods over rasterization for this body of work. Firstly, the computational savings promised by perceptual rendering shortens the cost gap between ray-based and rasterized methods. Perceptually based rendering with ray-based methods promises interactive framerates *and* physically accurate rendering; you can keep your cake and eat it too, you just won't notice you are not eating all of it. Secondly, when dealing with perception (especially in interactive virtual reality applications) it may be a substantial boon to keep the representation of light in the virtual world as close as possible to its physical counterpart. Not only is it more intuitive, it might better ease exploitation of perceptual phenomena based on light intensities, color, and other physical features. Finally, and just as importantly, it might motivate the development of mainstream dedicated hardware, such as commercial GPUs with integrated ray-specific calculations or widely available low-latency high-frequency eye tracking hardware.

2.6 Gaze-Contingent Rendering

Appreciating that the human visual system only perceives a small portion of its field of view in high detail opens the door to a technique more commonly known as foveated imaging, or foveation. Foveated imaging techniques take advantage of the low acuity in peripheral vision to reduce rendering quality at varying levels of detail, allocating more rendering time to fixation points corresponding to the fovea. Accounting for level-of-detail (LOD) with respect to *distance* is already commonplace in real-time graphics; MIP Mapping for textures being one such ubiquitous example. level-of-detail (LOD) based on peripheral vision, however, was relatively unused outside of academic research and specialised applications with large budgets [Tong and Fisher, 1984].

A steady reduction in cost and size of functional eye trackers suggests that foveated imaging may be commonplace in the future, especially in full-package peripherals like HMDs. There have been many studies conducted in order to investigate foveated imagery; the more recent of which are gaze-contingent systems that use eye trackers of some form. Foveation has been used for image compression, volumetric data observation, real-time graphics applications, and more recently in VR.

These applications can fall under two categories. Firstly, using sensory data and knowledge of sensory phenomena to produce **perceptually acceptable** content. An application that would fall under this category is perceptual static image and video compression; regions of lower visual saliency suffer the most compression, whereas regions with higher saliency are stored as close to source quality as possible. If a user were to look away from the intended or predicted scope, image quality would be noticeably poor. Secondly, the same knowledge can be used to produce **perceptually lossless** content, that is, lossy content which is perceptually indistinguishable from its lossless equivalent. An example would include adaptive rendering with eye tracking, rendering foveal and peripheral regions at higher and lower quality, respectively. The term was originally suggested by Duchowsky and Çöoltekin [Duchowski and Coltekin, 2007].

Perceptually lossless rendering implies dynamic content and gaze contingency, tracking and/or wholly predicting user fixations to ensure the user never observes a decimated region. Only a few studies in gaze-contingent rendering have fallen under this category, since limitations of tracking equipment have been overcome and underestimation of the problem has only been fully appreciated more recent times.

2.6.1 Window Boundary

One of the immediate decisions that come to mind with foveal imaging is whether the boundary between layers requires blending or can be left as a harsh transition. Reingold and Loschky [Reingold and Loschky, 2002] conducted a series of experiments with gaze-contingent displays, with their final experiment focusing on the effect of window blending. Their results showed that there was no significant performance difference (saccadic eye latency to a target) between harsh or blended window edges. More interestingly, they found that a fully low-pass image (all at peripheral quality) actually had higher mean performance (lower latency) than both gaze-contingent windowed imagery. An earlier study by Holmes et al. [Holmes et al., 1977] suggests that items in the direction of gaze interfere with the processing of other stimuli in the visual field. Similarly, the subjects in Reingold and Loschky's experiments may have suffered this effect. It should also be noted that the image degradation method used (simply a Gaussian blur) was subjectively noticeable and objectively reduced resolution well below the sensitivity limits of human vision in the periphery. Consequently, it would be imprudent to take these results at face value for perceptually lossless applications.

As far as we are aware, there have been no significant further studies focusing explicitly on the subjective and performance effect of window boundary types (sharp/harsh or smooth/blend) in gaze-contingent rendering. Further studies simply assume a blended

boundary, most likely due to intuition and the irregular sensitivity between participants that a blended boundary might alleviate.

2.6.2 Variable Resolution

Lower resolution rendering in peripheral vision is affordable due to the loss of acuity with increasing eccentricity. In some ways, it seems the most obvious method of saving computational load. Some of the earliest work on perceptual rendering used volume ray casting with variable resolution for volume ray casting. Levoy and Whitaker [Levoy and Whitaker, 1990] varied resolution as a function of the euclidean distance from the fovea's fixation point using discrete levels of detail, casting a ray for each pixel in the fovea, and one ray for more pixels in the peripheral regions. Although their system didn't conduct foveated rendering in real-time (all possible views of the image were pre-computed and then displayed accordingly), foveated renders took on average 1/5th of the time required for a full-resolution render. The subjects in this study were generally aware of the imagery's variable-resolution structure. The system displayed foveated imagery at 15 FPS with 150 ms latency, but it is not specified whether the authors believe this was due to system latency or excessive blurring in the periphery, but it was more than likely a combination of both factors.

Some more recent work on variable resolution foveation by Guenter et al. [Guenter et al., 2012] demonstrated actual resolution-based real-time foveation. The static scene in their experiment was rendered three times: once for the whole window at the lowest resolution; once for a large circular area at medium resolution, which roughly corresponds to the macula with a larger radius; and a final time for the smallest circular area, which corresponds to the fovea, at native resolution. Each rasterization was layered according to ascending pixel density and radially blended with the layer directly beneath it to ensure smooth window transitions. Additionally, the innermost layer (foveal) is rendered at 120 Hz while the two outermost layers are rendered on alternating frames (even and odd) at 60 Hz. Guenter et al. reported an increase in performance by a factor of 5-6, where subjects rated foveated sequences at an equal or better quality than the fully rendered counterpart.

With all variable resolution methods some sort of anti-aliasing is required to prevent distracting temporal artefacts in the periphery during motion. Guenter et al. combine three different methods (multi-sample anti-aliasing, temporal re-projection, and whole frame jitter sampling) to prevent this. They also recognised the importance of low system latency for effective foveation, specifying that the maximum acceptable system latency is much smaller than the 100-150ms estimate suggested by Levoy And Whitaker. For each frame, their implementation had a total system latency, start of

eye capture to pixel switch, between 23ms (best case) and 50ms (worst case). Subjects in this study still selected the non-foveated reference as the higher quality render in a pairwise comparison across all tested levels of foveation quality. In contrast, a ramped comparison, where foveation ramped upwards or downwards to/from reference, showed mixed results. In either case, the reference was only rendered at 40 Hz (as high as their system supported), and V-Sync was disabled for all experiments which may have affected the quality of the subjective results.

2.6.3 Refresh Rate Modulation

Smith et al. [Smith et al., 2014] foveate by way of refresh rate modulation, similar to frame-rate differences in peripheral and foveal regions in the work of Guenter et al. In contrast to spatial degradation, especially with regards to image resolution, refresh rate modulation does not require a costly blurring effect to compensate for aliased imagery. The ray-tracer featured in this works updates every pixel in the foveal region each render call and pixels in the peripheral region once every N foveal frames. Peripheral pixels are subdivided into m work groups, such that one and only one group is rendered per foveal frame and each group renders once and only once within the N frame window. Thus, each pixel in a particular group still only updates once every N frames. Rendering a Whitted scene with N equaling 12, Smith et al. achieved an increase in speed by a factor of 3.2 without secondary rays, 6.3 with secondary rays, and 2.8 for high-polygon scenery, representing a speed-up of more than 100 FPS. The value of N was chosen as it was the highest value reported where refresh rate modulation went unnoticed.

2.6.4 Contrast Sensitivity

Murphy et al. [Murphy et al., 2009] present a foveation method based on CSF (Section 2.2.2). Their method, using ray casting, conforms ray distribution according to the respective angular frequency, thus allowing image degradation without manipulation of the underlying scene geometry. This is done by casting rays for every pixel of a ray mask, which is the basis for an intermediate mesh composed of quads of increasing size as a function of their eccentricity from the point of reference (a discretised form of the contrast sensitivity function). When an intersection occurs the information is stored in the respective quad of the intermediate mesh. When a quad contains a heterogeneous mix of intersected and non-intersecting data, it is classified as an edge and are rendered by casting rays through all contained pixels.

Although they share some similarities, the crucial difference from the methods dis-

cussed in the prior two sections is that Murphy et al. are specifically informed by the CSF. As a consequence, and to maintain simplicity, the subject was kept at a fixed distance from the screen. They evaluated their method against the same method without the edge refinement described above and the reference. For a human expression location task, search time decreased as the full detail window increased, with lower search times with edge refinement. However, in contrast to the control group, both refined and non-refined modulations produced a lower mean number of fixations for all tested window sizes. Additionally, the mean accuracy from the control trials was lower than the accuracy on both other methods across all tested window sizes. Murphy et al. report an average performance inflection point or peak with a foveal region corresponding to 10°.

These results provide little guidance towards effective perceptually lossless acceleration, but do provide some interesting insights on how prominently task performance and aesthetic evaluation diverge. In a subsequent evaluation, supporting by findings from Cave and Bichot [Cave and Bichot, 1999], their control group claimed there were too many distracting faces in the search space, resulting in lower task performance. In contrast, the foveal methods produced perceptible changes in quality in the periphery and, interestingly enough, actually ran at lower frame rates than their control method ("just under 40 FPS" versus approximately 20 FPS for their foveated methods).

2.6.5 Detail Elision

The methods described so far target the final render, but not the content to be rendered. In contrast, the subsequent methods will do the opposite. Detail elision in particular focuses on polygonal mesh decimation. Obviously, this is not a permanent process (at least not for dynamic content) so decimation must be calculated on the fly. There are two ways of achieving this; pre-computation of the object at various levels of detail and switching between each at run time or multi-resolution/progressive meshes [Hoppe, 1996] which adjust contextually at run time. The former suffers from noticeable popping when switching between levels and a considerable amount of pre-computation. The latter, although providing smooth composition across levels, suffers from run-time overhead.

Oshima et al. [Ohshima et al., 1996] adopt the first method in their study on gaze-contingent adaptive rendering. They constructed hierarchical models for the objects in their scene at six increasing levels of detail. They use the observer's head position as an estimate for their gaze direction and do not update the image during saccadic movement, described within this study at $180\,^{\circ}\,\mathrm{s}^{-1}$. Their subjects reported that, when gazing at the fixation cursor, they only recognized a slight loss of peripheral quality.

The rendering rate increased from 4.5 FPS for a full high level of detail render to 20 FPS for an adaptive render. There is no mention of popping or its detrimental effect on the perception of the scene.

2.6.6 Simplified Physics

Another gaze-contingent method relies on simplifying collision handling and physics mechanisms outside the foveal region. An in-depth study conducted by O'Sullivan et al. [O'Sullivan and Dingliana, 2001] provides good insight on its application. Preliminary experiments with simple 2D collision of circular objects found an inability to detect gaps between collisions as eccentricity of the collision increased, with larger gaps being detectable for larger extents in the periphery. Additionally, with the addition of homogeneous/heterogeneous stimuli, performance decreased in the presence of homogeneous stimuli. Their 3D collision was based on a hierarchical sphere tree, processing collisions at a coarse level and successively refining based on the amount of time left for the frame. The 3D experiments were evaluated with free-viewing, unlike the 2D experiments where participants were asked to fixate such that the collision occurred peripherally. Interestingly, as the velocity of the collisions increased the perceived quality of complex physics simulations decreased. This suggests that at high velocities, it may be acceptable to reduce physics complexity at all locations in vision, and possibly even further so in the periphery. This agrees with prior findings by Profitt and Gilden [Proffitt and Gilden, 1989] that found that subjects only used one dimension of information with dynamical judgements, suggesting a confusion factor brought on overloading multi-dimensional information.

2.6.7 Saliency

As mentioned in Section 2.2.3, there is diverging opinion whether saliency is primarily bottom-up (salient features independent of task) or top-down (salient features dependent on task), but it seems more likely that is instead a combination of both with bottom-up dominating initial stimulus exposure (up to around 3 s) followed by top-down features. Exploiting scene saliency requires careful engineering, either to isolate and control a single aspect or ensure proper coverage of all feature fixations.

Typically, contextual saliency has been applied to image compression techniques and rendering priority optimization [Itti, 2004, Lee et al., 2002, Baudisch et al., 2003]. As mentioned, these are not typically intended for perceptually lossless viewing. For these techniques, it can be understood that there are four different approaches to perceptual encoding [Li et al., 2011]: firstly, saliency data (that is provided passively by

eye tracking or actively by user selection) can be passed to the encoder for transmission (see [Lee et al., 1999] for an early example); secondly, exploiting bottom-up attentional theory for contextually salient features by using machine vision to detect locations of typical regions of interest such as human faces and encoding those at higher qualities; thirdly, exploiting results from visual psychophysics literature, such as contrast sensitivity and noise tolerance, to degrade the image to a just-noticeable amount; finally, using top-down attentional theory we can predict gaze direction based on user goals, which is particularly useful when goals are induced such as task performance trials.

Borji et al. [Borji and Itti, 2014] have provided evidence that fixations convey information regarding the observer's mental state and task. Additionally, in some scenarios (such as video games) goal-driven feature saliency is easier to exploit given that goals are often made explicit as part of the design. Peters and Itti [Peters and Itti, 2007] observed top-down (eye tracked data and image regions) and bottom-up (spatial and temporal image features) gaze data for several game events and found that they were able to predict certain user-initiated game events based on fixation patterns leading up to the event.

Cater et al. [Cater et al., 2002] conducted experiments on the imperceptibility of foveal rendering using a top-down saliency model. Users were instructed to perform a simple counting task on a group of objects that moved within a pre-rendered scene, while a control group was told to freely observe the pre-rendered animations. Peripheral degradation in the ray-traced scene was achieved through a lack of anti-aliasing, motion blur, and allowance of a single light bounce. Each animation in a quality pairing combination was played in sequence during the trials. In both cases where the fully high quality render was shown before or after the peripherally degraded animation, almost all users performing the task (bar one) did not detect a quality difference while almost all users in the control group (free viewing) did. Additionally, in pairings where a fully high quality render was shown before or after a fully low quality render, approximately a quarter of the task-engaged users failed to notice a quality difference. Later results by Sundstedt et al. [Sundstedt et al., 2004] that also adopted a top-down saliency model support this study. Similarly to the O'Sullivan et al. experiments, these results agree with the Profitt and Gilden study.

Successful gaze prediction based on saliency therefore seems highly dependent on goal-oriented objectives. Outside of controlled experimental conditions, it seems very unlikely that we can rely on saliency alone to produce a perceptually lossless system. Indeed, Tatler et al. [Tatler et al., 2011] argue that the basic assumptions that underlie many of these studies are based on static and controlled viewing and shouldn't be generalized to how gaze is allocated during natural behaviour. Yet they suggest there

is a consistent set of principles (based on task, environment, and prior factors) in studies on natural gaze allocation that should be focused on instead. Regardless, it seems apparent that its use in combination with other methods may compensate for failures in other areas of the model avoiding, for example, the "pop" effect caused by latency-heavy tracking).

2.6.8 Color Degradation

Experiments by Watson et al. show that search performance is not greatly affected by colour degradation in the periphery compared to full colour, although the amount of degradation possible without significantly affecting performance depends on the difficulty of the search task. [Watson et al., 1997] There is no indication of their system being perceptually lossless, but given their focus on task performance it seems reasonable to say it was not.

As Duchowsky and Çöoltekin reported previously, there is very little work on peripheral colour sensitivity. [Duchowski and Coltekin, 2007] Although seven years have passed since their publication, this statement still holds true. In fact, to the best of my knowledge, there is absolutely no work done in the field of perceptually lossless colour degradation.

2.6.9 Foveated Image Quality Metrics

Traditionally, image quality metrics assume uniform quality perception at the foveal level across the entire image. Well known perceptually informed metrics such as Structural Similarity Index (SSIM) [Wang et al., 2004] and more recently HDR Visual Difference Predictor (HDR-VDP2) [Mantiuk et al., 2011] perform significantly better than other existing metrics for those scenarios. However, foveated imagery (particularly in rendering) is meant to be appreciated at a single point in space and time, and are not meant to be appreciated entirely at foveal fidelity but instead at the varying level of fidelity across the visual field.

There are a few examples of foveated image quality metrics. Wang et al. [Wang et al., 2001] introduce the FWQI, and they too note that most image quality metrics are designed for uniform quality images and do not correlate well to perceived quality at a single point in time. Lee et al. [Lee et al., 2002] introduce FSNR with moderate results, however PSNR (which the model extends) is simply a cumulative error metric with no perceptual information. Rimac et al. [Rimac-Drlje et al., 2010] introduce an extension to SSIM named FA-SSIM which outperformed the base metric on a video database simulating networking artefacts, but their method relies on temporal infor-

mation. Tsai and Liu [Tsai and Liu, 2014] introduce their own window-based foveated implementation of Structural Similarity Index (SSIM) using image saliency. Similarly, they claim higher performance on tested databases, but their method relies on the selection of an appropriate saliency model.

2.7 Health and Safety

Simulator sickness is an inherent consideration of all developing VR systems. There are a number of contributing conditions (and many more variables) that may influence its severity [Kolasinski, 1995]. One of the most widely accepted theories is the cue conflict theory, in which mismatches between expected movement and actual or simulated movement (due to errors or insufficiencies of the simulator) cause incidences of sickness. However, there have been reported situations with heavy cue conflicts in which subjects did not report any simulator sickness. In contrast, postural theory argues that there is some sensory redundancy and that the actual major contributor to simulator sickness is postural disequilibrium (ataxia). The many contributing factors make the phenomenon a very complex problem to tackle in its entirety. This work will address incidences of simulator sickness in a reactive manner, but some basic considerations will be taken into account prior to development.

Of course there are also additional concerns when illuminating the eye (and consequently the retina) with NIR light. The visual stimulus of the eye in response to NIR light is very low [Jaeger, 2009]. Consequently, the operations of the eye that normally regulate the amount of light hitting the retina (contraction of the pupil, gaze aversion, etc.) are not triggered. Existing safety standards define exposure limits for the retina, cornea, and skin[4] with the retina being the most susceptible to damage. Through the course of developing our own eye tracking system, special care has been taken to ensure the NIR illumination system falls within safety guidelines.

2.8 Literature Summary

As prior literature has shown, there have been several efforts towards the integration of visual perception in computer graphics. Additionally, this review has also presented a sample of existing studies into user-responsive hardware and the problems inherent to particular VRE systems (e.g. display panel versus projection based). It has also presented a number of existing gaze-contingent rendering techniques that target various parts of typical real-time rendering pipelines. What is sought to be accomplished by

[4]IEC-62471 (https://webstore.iec.ch/publication/7076)

this research is the integration and exploration of all of these fields under a cost-effective single project.

An advantage over prior research efforts that this body of work benefits from is the current state-of-the-art in hardware technology. In many cases, the study of particular methods or the exploration of complicated dynamic hardware systems was simply either not feasible or worthwhile given the cost. This work introduces cost-effective (albeit still at the enthusiast level) VREs (Section 5.4 & 5.3), high-performance eye tracking for general applications (Section 5.5), and cheap solutions for universal problems in VR (Section 5.1). The work also addresses closer integration of visual perception in existing hardware systems (Section 3.1) that may benefit tremendously from it and introduce new possibilities for use.

The hardware readiness level leads us to another question which hasn't been addressed in depth; maintaining perceptual losslessness of gaze-contingent systems. In particular, methods to address system latency across all parts of the VRE pipeline. This includes the discussion of methods to compensate for latency, to reduce rendering time in significant ways, and methods to accelerate the rapid prototyping of novel perceptually lossless foveated rendering methods. This research addresses those gaps by providing corrective measures for estimated system latencies (Section 4.1), the presentation of gaze-contingent rendering methods that target computationally costly aspects of the rendering pipeline, and the introduction of new image quality metrics specifically for foveated imagery (Section 4.2).

Chapter 3

Perception in Novel Scenarios

3.1 Digital Signage

As part of my work with Disney Research, I worked on an internal project with direct application to one of Disney's largest business interests. This work built on the knowledge of display technology gathered from the literature discussed in Section 2.4. In theory, it is possible to emulate a physical surface and its reflectance properties with an appropriately constructed electronic display technology. We can create "magical" versions of these reference surfaces, that can be directly manipulated like any other electronic display. Unfortunately, what makes a simulation believable is its robustness and accuracy under changing lighting and environmental conditions. Thus, although a replica may look believable under very controlled conditions, it will look incorrect the moment we break those conditions. The properties of emissive technologies and details of even "relatively simple" physical surfaces can be hard to account for.

Building on this, I provided a prototypical solution that could address a subset of these issues (and with further refinement, address many more) by capturing useful information from both reference (source) and replicating (target) surfaces. This proof in concept could capture the difference in color between a source pattern and the same pattern displayed on the target surface, from changing viewing angles (a limiting aspect of display technologies) and lighting conditions. From that captured color difference, the displayed pattern could be iteratively modified until it's emissive color matched the reflective color of the source material.

This system was built using a Raspberry Pi that was directly connected to the target display surface. A pre-determined pattern would then be displayed on the target. The source surface was placed within view of the camera as well, showing the same pattern (in this case, a Macbeth ColorChecker Chart). See Figure 3-1 for an image of the

system in action. Fiducial markers were used to ease tracking of the target and source surfaces across varying viewing angles.

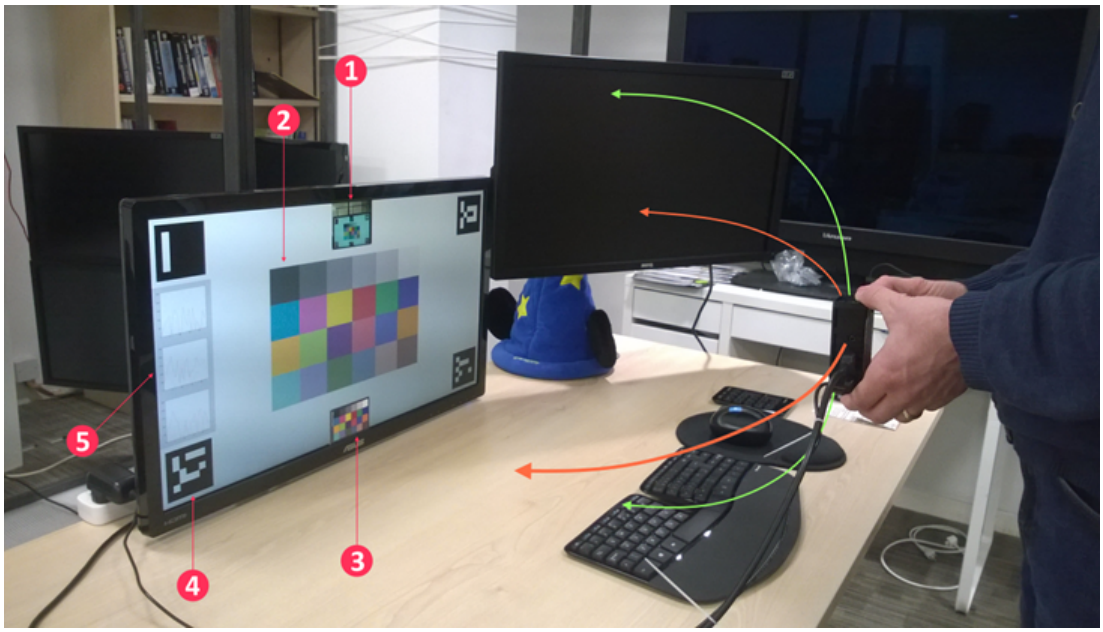

Figure 3-1: The prototype adaptive color system with (1) camera feed coming from the device to assist positioning, (2) the digital version of the test pattern, (3) the physical version of the test pattern on the reference material, (4) fiducial markers to assist tracking, and (5) plots showing the color differences between the captured reference and target colors.

The idealized version of this system would eventually use higher specification hardware, such as hyperspectral cameras for accurate color capture and comparisson, and have further capabilities to encompass other features of natural surfaces (e.g. specular highlight simulation by detecting position of incoming light sources). An interesting aspect of this work is that it is not solely dependent on uniform qualities of the display. The luminance profiles of displays can lead to poorer contrast or less brightness when viewing screens from eccentric angles. The viewing bounds, or viewing cone, is the frustum in which a display demonstrates acceptable visual performance.

An example of this can be seen in Figure 3-2. In these cases, the eye is located at a position perpendicular to the display surface. However, if the eye were at some angular eccentricity (say 45° away from perpendicular) while still observing the same point on the screen, the apparent colour profile of that particular point would be significantly different (and less bright), which is not the case in the reference surface.

This lends to a very interesting system requiring a vast number of environmental and user tracking capabilities. Additionally, how believable these simulated surfaces

need to be in order to convince the average user (is "close enough" enough?) would require its own in-depth study. My work on this project has resulted in my inclusion in a patent addressing the potential for such technology and a variety of sample use-cases. The project continues in the hands of other members of the Disney Research team.

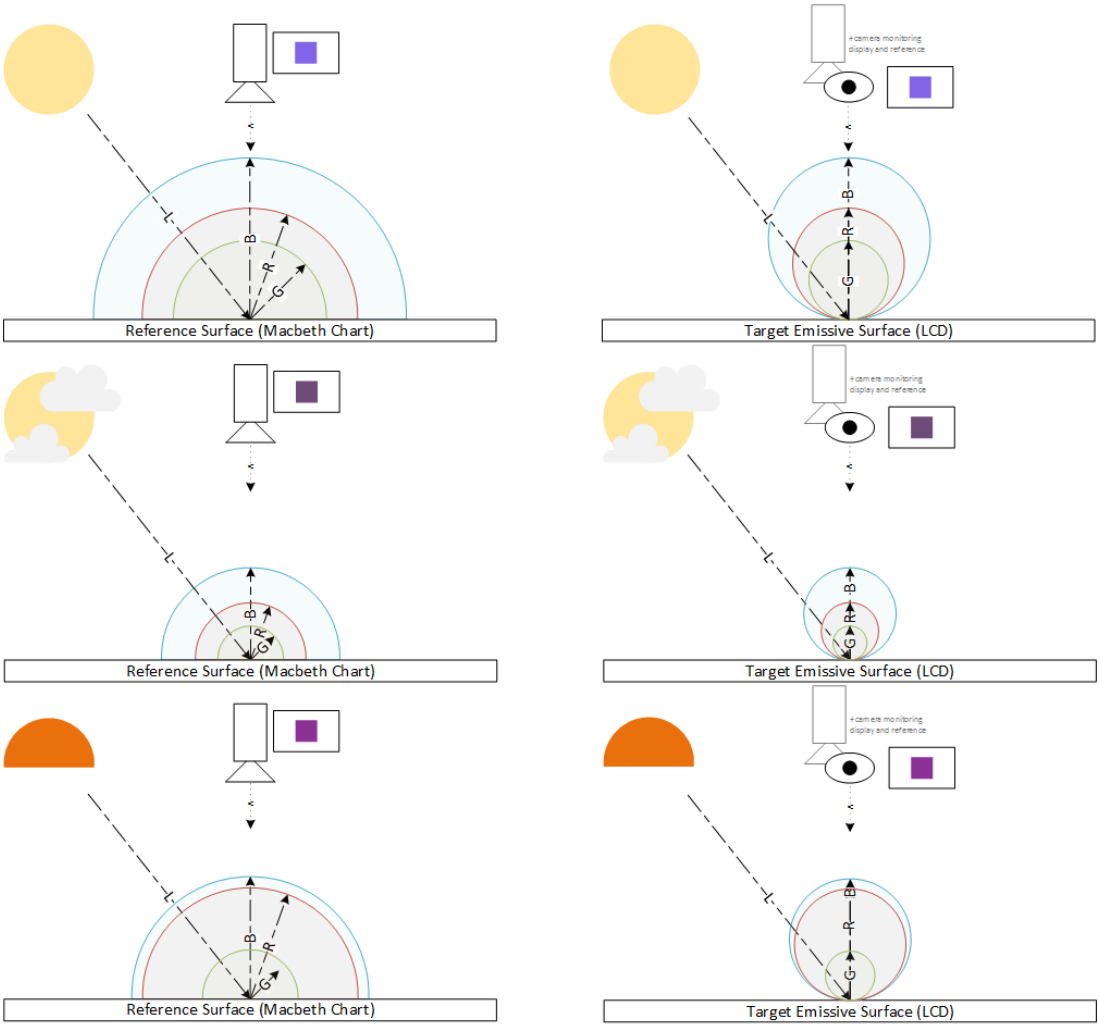

Figure 3-2: Adapting display light emissions to closely match existing physical surfaces. The left column shows the reflective properties of a given reference surface (in this case, perfectly diffuse) subject to different lighting conditions, divided into R, G, and B wavelength intervals. A camera captures the color reflectance from the reference which is used in the replication stage. The right column are the approximate replications of the reference surfaces using digital display technologies. Since the light emission profile for the display technology can differ in shape and size, the user's position should be used in conjunction with the captured reference color to approximate the emissive response on the display so that it matches as closely as possible to the reflective response from the reference surface.

Chapter 4

Lossless Perceptual Rendering

4.1 Latency Aware Foveated Rendering

My work in gaze-contingent rendering led to the implementation of a simple foveation system within Unreal Engine 4 (unreal), drawing elements from the resolution based methods discussed in Section 2.6.2. Unreal Engine 4 (unreal) is a widely-adopted commercial video game engine produced by Epic Games, Inc. The engine was chosen for its maturity and the free access to its source code (in contrast to other common commercial engines like Unity which require a premium to access source).

Although Parkhurst and Niebur [Parkhurst and Niebur, 2004] had previously implemented foveated geometric complexity decimation on a now decade old version of the Unreal Engine, their study study focused on task performance under aggressive foveation. For this implementation, focus was on addressing latency critical aspects of gaze-contingent rendering systems (discussed in depth by Guenter et al. and other work throughout 2.6). The research work was consolidated into a short research paper that was published in the ACM Proceedings for the 12th European Conference on Visual Media Production (CVMP 2015).

One of the primary objectives for gaze-contingent rendering is to prevent the user's awareness that any decimation is ocurring in the periphery. Universally, latency is a critical bottleneck for these systems, as they can produce a "pop" effect as the foveated render catches-up to the user's gaze location. There is no single source for this latency. It is produced in part by transmission latency in hardware (propagation of data across the system, individually minimal but collectively substantial), the delay in the eye tracking module (detection, correction, translation, and transmission), and the delay to render the actual scene (which can vary depending on scene contents).

Other virtual-reality systems (specifically, any that use real-time user-tracking data

streams for rendering, which now applies to any worthwhile system) suffer similar problems. Oculus, for example, use their TimeWarp feature to warp head-mounted display renders to match the most up-to-date position and orientation data prior to display. Without the TimeWarp feature many users experience nausea with rapid movements as the HMD "swims" across the scene. TimeWarp is not a silver bullet, however, as it can introduce disocclusions which must be addressed appropriately.

Similarly, this work introduces a simple formulation (see Equation 4.1) that takes an estimated maximal total system latency, the user's distance from the screen, and user's maximal saccadic speed (which depends on expected range of movement) to increase the foveal render size dynamically. The consequence is that render time is increased, but we ensure that there is never a "pop" effect as the user's foveal field-of-view is always constrained to a foveal-quality area of the render.

$$F_{\varnothing} = 2\,\rho_{pixel}\,d_u \tan\left(L_{tot}S_{max} + \frac{\alpha}{2}\right) + 2\,b_w + c \tag{4.1}$$

Equation 4.1 is relatively straightforward; L_{tot} is the worst-case total tracking latency in milliseconds, S_{max} is the maximum saccadic speed in $radians/ms$, d_u is the user's distance from the screen, α is the angle subtended by the fovea which is roughly 5° [Osterberg, 1935], b_w is the width of blending border, and ρ_{pixel} is the pixel density of the screen in $pixels/mm$.

The error constant c is added due to some simplifications. We assume the user remains at a constant distance from the screen between each tracking frame. As the user's maximal positioning speed is unknown, we either employ a conservative position prediction model or reduce the total tracking latency until the difference of distances is near zero. We also avoid calculating the change in radius when gaze is not perpendicular to the display surface, as this depends on the angle of incidence and the curvature of the display. Lastly, we assume the tracker's precision and accuracy errors are negligible.

Although determining the user's distance from the screen is not trivial, the commercial eye-tracker used for this work provids via its API. Prior work shows that saccadic speed is positively related to the breadth of movement, namely that broader angular movements are also faster [Abrams et al., 1989]. Based on the size of the display and an estimate volume for typical head movement, this corresponded roughly to a maximal saccadic speed of $200\,^\circ\,\mathrm{s}^{-1}$. Maximal system latency was set to an estimate value for a average-case run through the system. The implementation was not dynamic to system latency due to its downside; increasing the foveal window increases the total latency which lends to a further increase in size, leading to a feedback loop. Addressing this issue appropriately is a matter of further research.

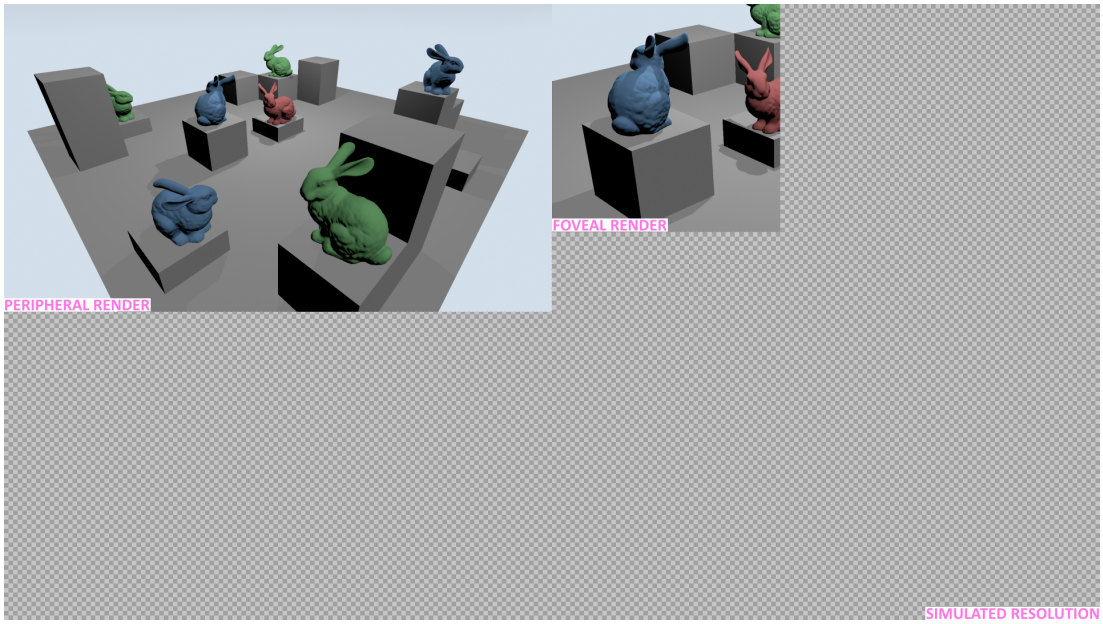

Figure 4-1: View of a foveated resolution render with moderate settings pre-composition, with relevant regions annotated. The checkerboard area represents the remaining amount of pixels that are required for a reference render of the resolution we are simulating. For example, using the same settings to simulate 4K UHD rendering, we would only have to render approximately 2.7 MP instead of 8.3 MP.

Although the rendering method is basic, it is an initial technique that can easily be exploited for perceptual losslessness. The peripheral render is rendered at a quarter of the intended display resolution and then upscaled with minor blurring. The foveal render, with diameter calculated by Equation 4.1, is layered on and blended against the peripheral layer (vignette mask) at the gaze point (see Figure 4-1).

A foveal diameter corresponding roughly to 1000 pixels within a 4K UHD render of a simple scene, displayed on a 28" monitor at a typical viewing distance of 20", was sufficient to reasonably compensate for total system latency representing a saving of approximately 6.8 MP. Note that with a near-zero total latency, the foveal window would only have to be approximately a third of that diameter. There are no official statistics on the overall latency of the tracker, but as the simple scene ran at a relatively high frame rate (approximately 60 FPS) the tracker was the most likely source of total system latency. Unfortunately, the total system latency when rendering at 4K UHD for our complex scene was too high to avoid the "pop" effect with a reasonably sized foveal diameter. Using the foveal diameter derived from our simple scene, our complex scene ran at approximately 24.3 FPS with foveation on compared to approximately 14.0 FPS with foveation off, an average saving of 20 ms per frame.

Figure 4-2: Comparison stills of the foveated rendering implementation within Unreal Engine 4 on the Elemental Tech Demo provided by Epic Games, Inc. The peripheral layer is upscaled from a lower resolution render and the foveal layer rendered at native density. On the left, the foveal region is focused on the character. On the right, the foveal region is focused on the lava fall behind him.

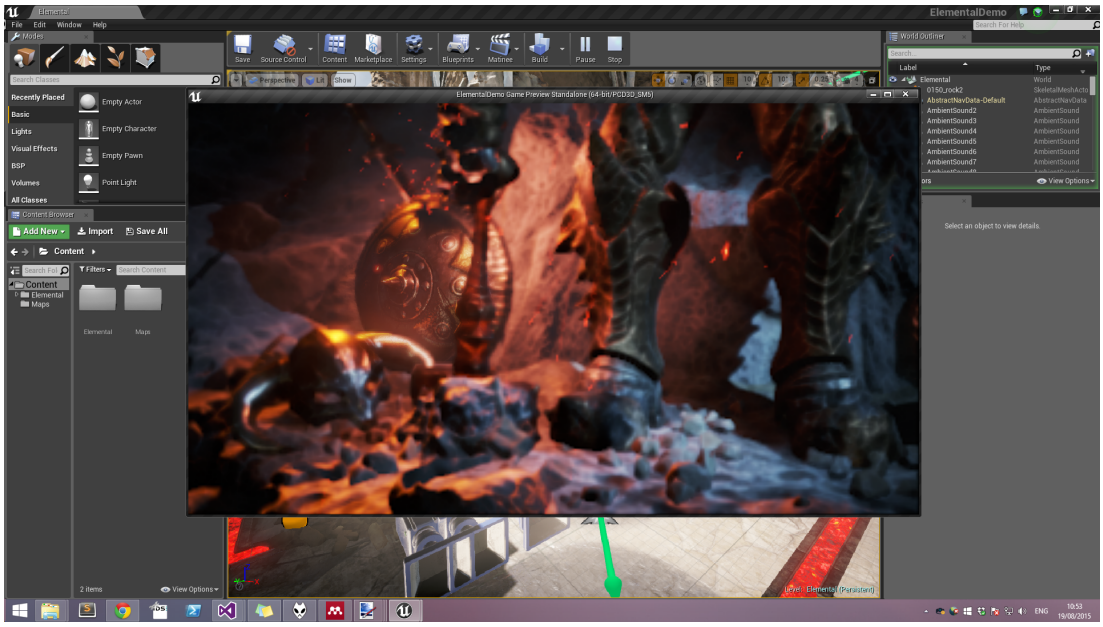

Figure 4-3: Full screen still showing the changes working within the Unreal Engine 4 editor environment. The foveal region is focused on the character's shield.

This study was mostly concerned on addressing latency, a somewhat overlooked

concern with foveated rendering methods. However, much like refresh rates and pixel persistence with displays for HMDs, ensuring low latency in eye tracked system is a balance between cost, constraints, and quality. In most cases this is a hardware engineering and system design problem, but this work sought to show ways to address the issue via software.

An alternative to compensation (subject of this paper), we can employ predictive methods. This is mentioned somewhat briefly in the paper, as accurate estimates for total system latency are also dependent on rendering time. A naive implementation would cause a feedback loop, where foveal window size increases, increasing latency (through render time), increasing size, ad infinitum. Implementing a machine learning method that is either predictive of user gaze (complicated, as indicated in Section **??**) or intelligent enough to balance latency versus quality would be necessary.

Irrespectively, the primary concern of this study was the lack of controlled user validation of the method (although the base method itself is not novel, just the implementation). This was later addressed (albeit only on the spatial domain) by further work on developing perceptually lossless methods and enhancements, discussed in Section 4.2.

4.2 Evaluation of Foveated Rendering Methods

I conducted further research with Disney Research on the method discussed in Section 4.1 as well as developing three additional methods inspired from prior work in the field. The three remaining methods target tessellation, Screen-Space Ambient Occlusion (SSAO), and reduced sampling for real-time ray-based rendering.

The methods are evaluated for their detectability at various parameter levels against real users, with the exception of the ray-based method which is evaluated against an extension developed for the HDR-VDP2 metric. Part of our aim in this study was to determine the adequate quality settings for each method that maintained perceptual losslessness. All four of our foveated rendering methods operate in real-time in their respective frameworks.

4.2.1 Screen-Space Ambient Occlusion

Ambient occlusion [Pharr and Green, 2004] is a well known technique in graphics to simulate the effect on diffuse lighting caused by occlusions created by objects present in the scene, including self-occlusions. It has been adopted to simulate a diffuse term that supports a complex distribution of incident light. Because ambient occlusion can be quite expensive to compute in real-time for dynamic scenarios, screen-space approaches are currently widely popular [Bavoil and Sainz, 2008].

We exploit Screen-Space Ambient Occlusion (SSAO) by varying the number of per-pixel depth-buffer samples in the foveal and peripheral fields of view. Although a very low number of per-pixel samples can cause banding (see Figure 4-4), we expect these differences to go unnoticed in the periphery due to the loss of visual acuity and contrast sensitivity. The scene we chose for this method is the Sibenik Cathedral populated with Stanford bunnies, as it provides a lot of occluding meshes with small details.

4.2.2 Terrain Tessellation

Our third method is a foveated implementation of a terrain renderer exploiting GPU-level tessellation. Geometry tessellation is a vertex processing stage that adaptively subdivides coarser geometry patches on-the-fly into smaller geometric primitives to generate nicer and smooth-looking details. Tessellation has been incorporated on modern GPU rasterization pipelines and is commonly driven by some view-dependent criteria. We chose this technique due to its wide adoption within the graphics industry.

Our foveated rendering method builds on an OpenGL framework exploiting tile-based tessellation. In order to determine the appropriate level of tessellation, we project the foveal window from screen coordinates into the scene. If a tile falls within either

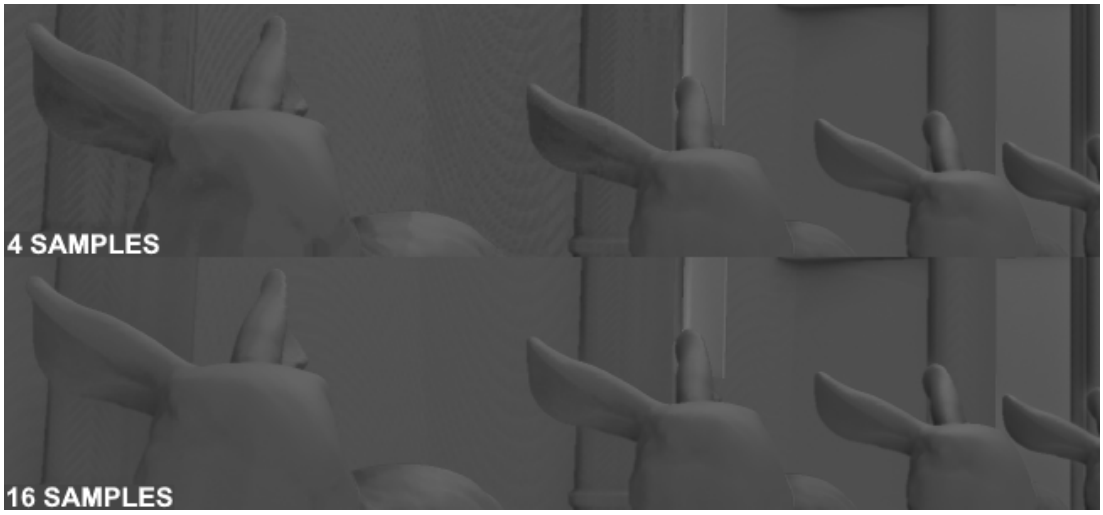

Figure 4-4: Strips from two foveated renders with the same fixation point but different peripheral sampling levels. Fixation point is at the bottom-right corner for each strip. Transition between foveal and peripheral regions are handled smoothly. At 4 samples there are noticeable artefacts, such as banding on the wall.

the foveal or peripheral field of view, the level of tessellation is set statically to the appropriate level. If the tile falls between the two regions (on the blending border) the level of tessellation is linearly interpolated between the two levels. Figure 4-5 provides a wireframe view with exaggerated settings of our method in action.

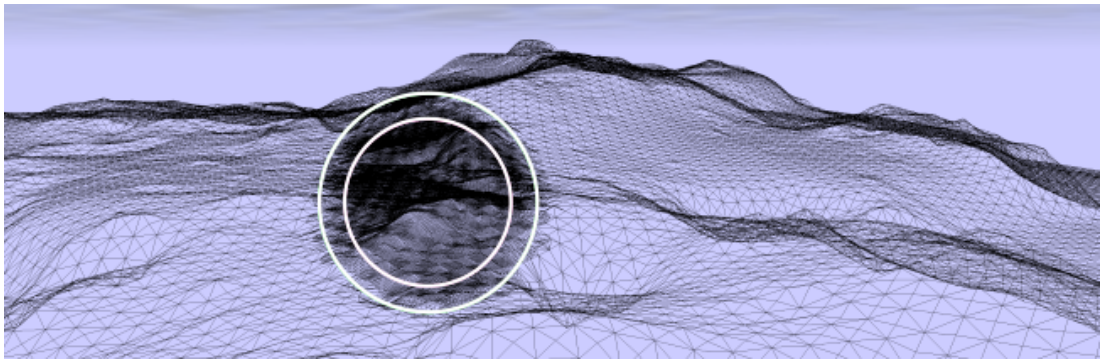

Figure 4-5: Wireframe view of a still from our foveated tessellation method. The foveal region is within the inner circle, the blending border between the inner and outer circles, and the peripheral region is outwith the outer circle.

4.2.3 Foveated Real-time Ray-Casting

Our fourth and final method, which we evaluate against the parametrized metric, employs foveally selective ray casting for 360° immersive virtual reality content, ren-

dered using a variant of multi-layer relief mapping originally developed by Policarpo and Oliveira [Policarpo and Oliveira, 2006], which allows motion parallax within a limited envelope of movement. The method normally casts rays to geometry and detects intersections with a given number of depth layers, represented as a series of RGBA textures mapped on the geometry. We vary the number of per-pixel ray-casting steps across the field of view. This can cause significant dis-occlusion errors and stair-stepping artefacts if the number of steps is too low. Again, building on the lowered contrast sensitivity and visual acuity in peripheral vision, we expect there will be a balance between the severity of dis-occlusion and the number of per-pixel stepped samples that is sufficiently unnoticeable yet yield high performance.

Figure 4-6: Top: Sample frame from our ray-casting method with 120 per-pixel steps in the foveal region (within circle) and 10 per-pixel steps in the peripheral region (outwith circle). Bottom: Close-up of right lamp showing artefacts across different quality levels.

4.2.4 Foveated Image Metric

We wish to develop a suitable image quality metric specifically for foveated imagery to assist with foveated rendering method evaluation in the future. User trials are typically time consuming and costly, so their use should be reserved for methods that have reasonably high chances of success. However, perceptually informed metrics that take foveation into account are relatively unexplored (see Section 2.6.9). Instead of adopting and/or altering one of the aforementioned foveated metrics, we present a new metric that builds on an existing algorithm demonstrating a strong psychophysical background but lacking consideration for loss of visual acuity with eccentricity.

To this end, we extend HDR Visual Difference Predictor (HDR-VDP2) as it has a strong perceptual background, reports relatively good performance, is freely available, and is well documented. In order to improve the algorithm meaningfully, we targeted the degradation of contrast sensitivity in peripheral vision. We introduce the Cortical Magnification Factor (CMF) to the algorithm, as it describes the cortical surface area dedicated per degree of visual field with eccentricity, as a theoretically motivated parameter to calculating the extent of peripheral degradation.

There is a strong relationship between the CMF and the degradation of contrast sensitivity and visual acuity with visual eccentricity [Virsu and Rovamo, 1979]. Difference between contrast sensitivity or visual acuity in central and peripheral vision could be accounted for by compensating stimulus size by the CMF. We scale the contrast sensitivity function by the CMF at a given pixel divided by the value of CMF at fixation. For HDR-VDP2, we target the neural contrast sensitivity function [Mantiuk et al., 2011] which discounts light scattering and luminance masking.

$$CSF_e^{\boldsymbol{M}} = CSF_e - CSF_e \times (1 - \frac{\boldsymbol{M}_e}{\boldsymbol{M}_0})^{1+\alpha*(1-S)} \tag{4.2}$$

Where e is an eccentricity corresponding to a pixel position (x, y), CSF_e is the Contrast Sensitivity Function at that eccentricity, $\boldsymbol{M}_e$ is the CMF at that position, and $\boldsymbol{M}_0$ is the CMF at centre of vision. As HDR-VDP2 uses a multi-scale decomposition process, we increase sensitivity of detected contrast as scale decreases (S being 0.5, 0.25, etc) to allow the model to remain sensitive to large scale contrast changes over the visual field. Finally, α is a tunable parameter that we introduce to attenuate the effect of peripheral sensitivity.

4.2.5 Hypotheses

How perceptually lossless a foveated render appears to be can be determined by how reliably an average user would be able to distinguish the reference render as the

higher quality render when also presented to the foveated render. Thus, to validate our methods and determine whether they are perceptually lossless, the average user should identify the reference render (uniformly high quality) over the foveal render (high quality window at fixation, lower quality elsewhere) worse then chance. The more significantly different from chance this value is, the more reliable is the foveated method/quality pairing. We advance the following hypotheses, such that when comparing a reference and a foveated render:

H_0 *The average viewer identifies the reference render as the high quality render at chance ($\approx 50\%$ of the time).*

H_1 *The average viewer identifies the reference render as the high quality render better than chance ($> 50\%$ of the time).*

H_2 *The average viewer identifies the reference render as the high quality render worse than chance ($< 50\%$ of the time).*

H_2 is our preferred hypothesis, as it indicates the reference render cannot reliably be identified as the higher quality render. A failure to reject the null hypothesis does not allow us to make any conclusions on the effectiveness of the method. If results favour H_1, the method/quality pairing must be abandoned as the difference is reliably detectable.

4.2.6 Rendering Parameters

We use Equation 4.1 to calculate the foveal window size for our study. The fovea subtends the central 5° of radial area on the retina [Polyak, 1941], however we increase the value used in our studies to 9° to encompass the parafoveal area (approximately 7° of eccentricity) and to account for tracker error. This corresponded to a foveal window diameter of approximately 588.4 px (given the information in Section 4.2.9), which we round up to 600 px to account for minor accidental gaze drift.

The blending border between both regions is an additional 100 px, which is decided arbitrarily. Prior studies have shown that blending, or lack thereof, provides no significant user performance difference [Reingold and Loschky, 2002]. However, the peripheral degradation in that study was noticeable and may have interfered with the results. As far as we are aware, there are no further studies that focus explicitly on this subject.

We select three levels of detail per method to experiment on and to ensure some coverage of the parameter space. These three levels of detail are described throughout

this section as **low**, **medium**, and **high**. Low settings were chosen to provide the largest computational gain, but the most likelihood of detection that could still justify foveation. Contrarily, high settings were chosen as very unlikely to be detected, but with the lowest computational gain that could still sufficiently justify the use of foveation. The medium setting was chosen as the middle point between the two, an intuitively ideal balance between likelihood of detection and performance. See Table 4.1 for exact values.

	Resolution (scaling)	SSAO (samples)	Tessellation (levels)
LOW	0.25	4	8
MED	0.50	16	16
HIGH	0.75	64	32
REF	1.00	128	64

Table 4.1: Peripheral quality parameter values used in our study. For the resolution method, we render the periphery at parameter value of the target resolution and then upscale. For the ambient occlusion method we vary the number of samples. For tessellation, we vary the refinement of the tessellated grid per tile.

4.2.7 Fixations

For our experiments we decided to focus exclusively on perceivable spatial artefacts for our methods. Although we understand the importance of evaluating our methods temporally, our work serves as a preliminary study in automated and subjective evaluation of gaze-contingent methods. As our extension to the HDR-VDP2 metric (and the base metric itself) does not take temporal factors of human vision into account, we would be unable to accurately evaluate the perceptibility of our modifications through the image quality metric in a temporal setting. Additionally, due to the tracking hardware available to us (see Section 4.2.9) we would not be able to isolate our experiments from external error, leading to potentially flawed conclusions about the methods' perceptibility. We instead adopt fixation-based testing and use our tracking hardware to validate fixations.

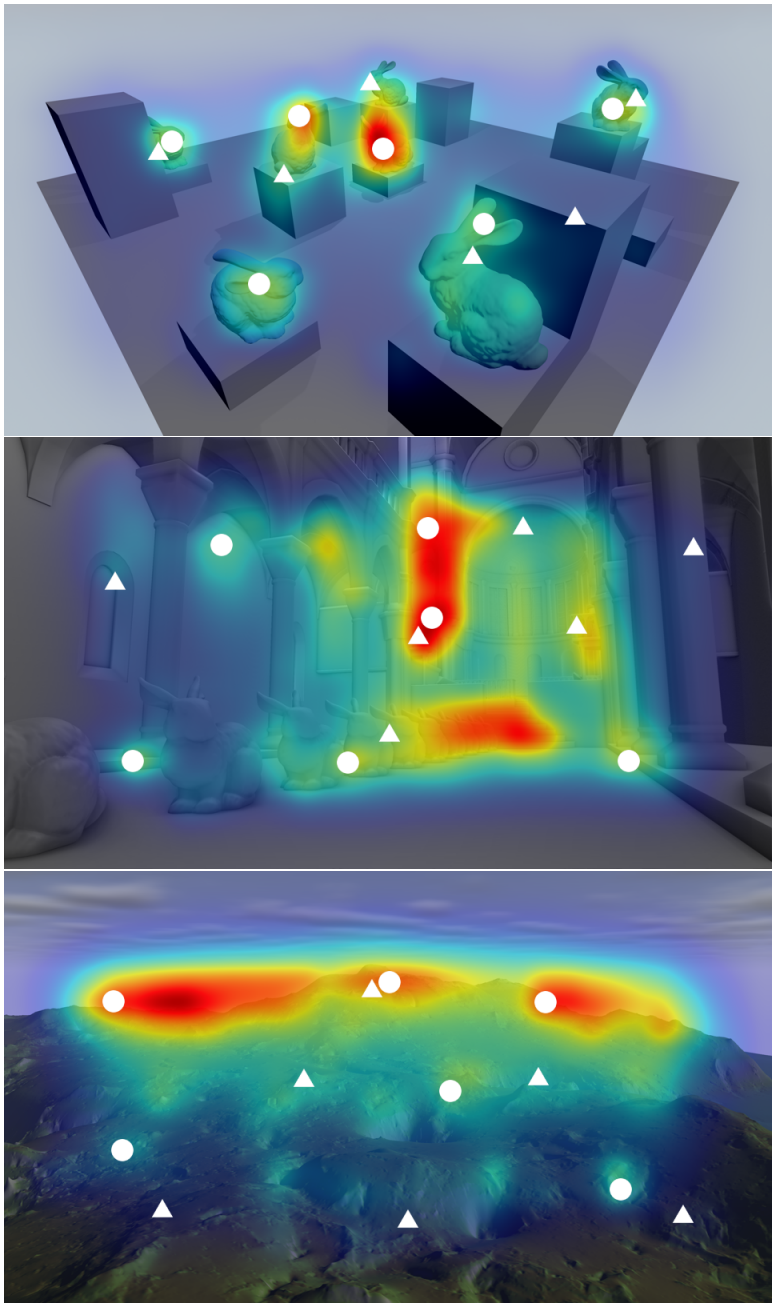

Figure 4-7: Reference renders for each method with respective Graph-Based Visual Saliency (GBVS) heat-map overlay, from top to bottom: resolution, SSAO, and tessellation. Circular marks denote fixations selected by GBVS. Triangular marks denote fixations that were selected subjectively by the authors.

Fixation-based testing introduces a few problems when evaluating methods for user preferences, image quality metric results, and reported computational load. In terms of computation, the position of the foveal render can greatly affect rendering times

depending on the method (e.g. tessellation on simple versus intricate surfaces). In terms of user preference, prior studies suggest that poor selection of the foveated region (such as random or brute-force selection) could lead to lower perceived image quality [Bailey et al., 2009]. In terms of image metrics, it must be general enough to provide realistic results for the phenomena it is modelling (in this case, the human visual system), where simplifications can lead to excessive positive or negative performance. Temporal testing does not suffer from these specific issues as gaze is a direct reflection of user preference and real-world data (which would validate averaging for computational results, for example).

In order to select plausible fixations we conducted a small pilot study, collecting gaze positions over a 10 second period during free-viewing sessions of our reference renders. We then ran Itti, Koch, and Neibur (ITTI) [Itti et al., 1998], GBVS [Harel et al., 2006], and Erdem and Erdem (CovSAL) [Erdem and Erdem, 2013] on our reference renders to select the saliency maps which fit closest to our collected free-viewing fixation data. The saliency model that most closely fit our data was GBVS, from which we select the centres of the 6 most salient, non-overlapping image regions. We also subjectively chose 6 additional fixation points which we found to demonstrate high detail variability or represented interesting regions of the image. The fixation points for each method/reference render can be seen in Figure 4-7.

4.2.8 User Trials

The experiments consisted of a number of tests in randomized order comparing a foveated render to the reference render. For each trial, a foveated render was displayed before or after a reference render for the same amount of time. Once both images had been displayed, the subject would then have to decide whether the first image appeared higher quality, the second image appeared higher quality, or if both images appeared identical, and respond appropriately.

Each subject underwent three test blocks, one for each rendering method, in randomized order. A test block consisted of 81 trials in randomized order. Out of these 81 possible trials, 9 were control trials while the remaining 72 were test trials. The amount of test trials are divided equally among each of the three quality levels, lending to 24 test trials per quality level per method. Of these 24, there are 2 trials for each of the 12 fixation points; one trial in which the foveated render is presented first and one where the foveated render is presented second. For the 9 control trials, 3 trials display the reference render against itself and 6 trials compare a fully peripheral quality render against the reference (per quality level and per first/second order).

The procedure for a single trial was as follows. Firstly, a neutral grey screen would

appear for two seconds. Then a small cross would appear on the grey screen indicating where the user was to maintain their fixation. Users were instructed to fixate at that position until the end of that specific trial. The eye tracker would ensure the user's gaze was fixated on the indicated area and would signal the start of the test. At this point, the first image in the trial would appear for two seconds, followed by the neutral grey screen with the cross at the same location for one second, followed by the second image in the trial for two seconds. If the user's gaze drifted away from the indicated fixation point at any time during the trial, the trial would not be interrupted but the results would be marked invalid. Finally, the neutral grey screen would return without the cross to await the user's response (first was better, second was better, or both appeared identical).

The user population consisted of 9 participants (1 female, mean population age of 32) who were computer graphics professionals with diverse backgrounds. All users had 20/20 or corrected to 20/20 vision. The eye tracker (see Section 4.2.9) was calibrated for each user individually before their testing session. Users were allowed to take short breaks at any point during a block (provided this was done at the answer screen for a trial and they remembered their answer) to avoid fatigue. Between each block, breaks of any desired length were allowed and users could leave the testing area, also to prevent fatigue.

4.2.9 Equipment

We use an Acer CB280HK 4K UHD monitor with a display area approximately $62\,\text{cm} \times 34.5\,\text{cm}$ in size, corresponding to an approximate pixel density of $6.23\,\text{px}\,\text{mm}^{-1}$. For eye-tracking, we used Tobii's EyeX commercial level eye tracker with 9-point calibration, with no accuracy and precision reports [1] and no specified latency at time of purchase [2], although internal testing yielded an approximate latency of 50 ms to 75 ms. Due to these specifications, we would be unable to reliably validate our methods temporally, and so our study focuses solely on spatial detectability. To easily accommodate the eye tracker's tracking volume and increase tracking accuracy, users were secured on a head-rest at a distance of 600 mm from the monitor for all experiments. For our rendering and benchmark tests, we use a desktop computer equipped with an Intel Core i7 4820K CPU and an ASUS R9 290X GPU.

[1] http://archive.is/qWvMi
[2] http://archive.is/o7b1M

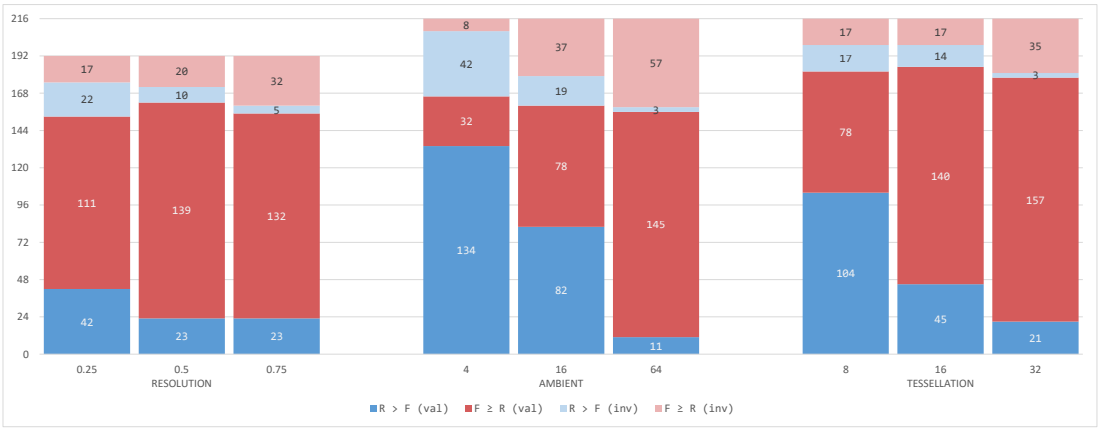

Figure 4-8: All trial results (excluding controls), split per method and per parameter setting. Valid instances where the reference was marked higher quality than the foveated render are in blue (invalid in light blue). Valid instances where the foveated render was marked equal or higher quality than the reference are in red (invalid in light red).

4.2.10 User Trial Results

All subjects completed all trials for all three blocks. However, one subject's resolution block trial data had to be discarded due to a misunderstanding of testing procedure, which led to all responses being invalid. This data was removed from our results and there were no other changes made to the data set. All significance values are evaluated at the $\alpha = 0.01$ level.

The proportion of invalid responses to valid responses was similar across parameters within a given method, with $\approx 18\%$ invalid responses for the resolution method and $\approx 16\%$ for the tessellation method. However, the ambient method demonstrated an overall higher proportion ($\approx 26\%$) of invalid responses when compared to the other two methods. Given that trial block order was randomized we exclude fatigue as a possible cause, and tracker error would have manifest itself across all trials. This suggests that the method may have caused distracting artefacts or the scene contained sufficiently distracting features to draw gaze. However, whether this difference is statistically significant is not determined.

Data for several quality/method settings demonstrate a "correct" (identified the reference render as the higher quality render of the pair) to "incorrect" (identified the foveated render as the higher quality render of the pair, or indicated that the quality of both were identical) ratio that was statistically significant in favour of H_2, thereby encouraging their adoption. These quality/method settings were **ambient high** ($P_{val} \approx 6.895 \times 10^{-9}$), **tessellation medium** ($P_{val} \approx 0.0011$), **tessellation**

high ($P_{val} \approx 4.598 \times 10^{-7}$), **resolution low** ($P_{val} \approx 0.0023$), **resolution medium** ($P_{val} \approx 7.938 \times 10^{-5}$), and **resolution high** ($P_{val} \approx 5.822 \times 10^{-5}$). The remaining quality/method settings either favour H_1 (ambient low), thereby discouraging their use, or fail to reject H_0 (tessellation low and ambient medium).

Subjective responses from users suggest difficulties in distinguishing the images for the resolution trial block, with some subjects asking whether they were being shown different images at all. The users added that there were a few "obviously rough looking" images that they felt were easily distinguishable. These were most likely the control trials and a subset of the low quality trials. Subjects also reported the most confidence after the ambient tests, stating that the quality difference for many of the trials was clearly distinguishable. For the tessellation trials, user confidence was mixed, but overall subjects believed that they had identified the reference correctly.

4.2.11 Quality Metric

Using the results from the user trials, we parametrize our metric. The metric will then be used to evaluate our fourth and final foveated rendering method for immersive content. We first determine the ideal parameters for base HDR-VDP2, namely the peak sensitivity of the metric (p_{sens}), the excitation (p_{mask}), and inhibition (q_{mask}) of the visual contrast masking model. These are the tunable parameters provided by the base HDR-VDP2 metric.

HDR-VDP2 predicts the probability that the differences between two images are visible to the average observer (with 0 indicating impossibility and 1 indicating absolute certainty). To compare against the model's predictions, we derive our predictions from the data by comparing metric results against user testing results for the fully peripheral quality versus reference control trials. In this way, the base parameters for the HDR-VDP2 metric are calibrated for degradations at foveal fidelity (highest fidelity in the visual field).

We were unable to find a single set of base parameters that provided detection probabilities close to our data for all three methods. Therefore, we provide parameters per method and evaluate our selective ray casting rendering model against each. For the resolution data we use $p_{sens} = 1.0$, $p_{mask} = 0.14$, and $q_{mask} = 0.19$. For SSAO we use $p_{sens} = 0.8$, $p_{mask} = 0.54$, and $q_{mask} = 1.50$. For tessellation we use $p_{sens} = 0.8$, $p_{mask} = 0.54$, and $q_{mask} = 0.30$.

We then calibrate our extended metric using the attenuation parameter α from Equation 4.2, using the V1 cortex parameters from [Dougherty et al., 2003] for the CMF function. The detection probabilities output by our metric are compared against the foveated detection probabilities from our data; the number of valid and correct

responses over the total number of valid responses. The attenuation values we found to have the best fit were $\alpha = 2.45$ for the resolution data, $\alpha = 4.45$ for the ambient data, and $\alpha = 0.43$ for the tessellation data. Using our metric, the average detection predictions per quality setting per method (averaged over all foveated images in that class) can be seen in Table 4.2.

	Resolution ($\alpha = 2.45$)	SSAO ($\alpha = 4.45$)	Tessellation ($\alpha = 0.43$)
LOW	0.32 (0.27)	0.88 (0.80)	0.65 (0.57)
MED	0.02 (0.14)	0.29 (0.51)	0.12 (0.24)
HIGH	0.01 (0.14)	0.08 (0.07)	0.01 (0.11)

Table 4.2: Average predicted detection probabilities per setting per method (averaged over all foveated images in that class) from our extended metric, with probability values extracted from our data shown in parentheses.

4.2.12 Immersive Motion Parallax Rendering

We run our fully calibrated metric on our fourth and final method. For this dataset, we adjust the equipment and set-up specific base parameters of HDR-VDP2 to match values for a typical modern and commercial head-mounted display. In our case, we use the Oculus Rift DK2's resolution, screen dimensions, and typical eye distance from the screen. Renders from this dataset are then evaluated with our metric using the three parameter sets (one per method) derived in Section 4.2.11. The detection probabilities returned by our metric on this dataset are found in Table 4.3. Similarly to the other foveated rendering methods, we are only evaluating the method spatially at a single point in time. In this case, we use a single fixation point (in this case the flower pot in the scene, see Figure 4-6) and evaluate over a wider quality parameter space.

	Res. Settings	SSAO Settings	Tes. Settings
10 steps	0.50	0.37	0.03
20 steps	0.14	0.08	0.01
40 steps	0.04	0.02	0.01
80 steps	0.03	0.01	0.01

Table 4.3: Predicted detection probabilities for our fourth foveated rendering method, with foveal region rendered at 120 steps and periphery rendered at step rate listed in first column.

Out of the three parameter sets, the tessellation parameters seem to provide the

most unrealistic results given the amount of degradation at lower steps. Since the artefacts produced by reduced peripheral resolution are similar to those produced by reduced sampling (loss of contour and texture fidelity, etc.) we use the resolution parameter set for our metric to determine the ideal balance between detectability and computational performance for this particular method in Section 4.2.13.

4.2.13 Performance Gains

To evaluate computational performance we settle on the lowest quality setting per method that favours H_2, run our methods in real-time at each fixation point, and average the render time over 1000 frames. After which, we average across all fixation point times per method to provide the average rendering time for our method overall. We select **resolution medium**, **ambient high**, and **tessellation medium** for our quality settings. We chose the resolution medium over resolution low in order to be conservative with our estimates, as detection probabilities appear to plateau between the two.

The average render time over all fixation points, the fixation point with the worse average render time, and the fixation point with the best average render time compared against the average render time for the reference *per method/quality setting* are show in Table 4.4. The table also includes the average rendering time for our foveated ray-casting method at the flower pot fixation point at the 20 step quality level.

	Optimal Settings	Reference
Resolution	7.18 ms (7.01 ms / 7.27 ms)	14.69 ms
SSAO	22.31 ms (21.17 ms / 25.2 ms)	82.34 ms
Tessellation	5.88 ms (4.54 ms / 10.16 ms)	17.24 ms
Sampling	19.61 ms	28.57 ms

Table 4.4: Mean frame rendering time over all fixations per method/quality setting in milliseconds. Fixations with the best and worst (respectively) mean render time shown in parentheses. Resolution, SSAO, and tessellation methods are targeting 4K UHD while the Sampling method is targeting 1600×1018.

4.2.14 Analysis

Overall, all of our methods enjoyed some success. As expected, the low quality settings were the most easily detectable, but with the resolution method the difference between settings was much less substantial than initially expected. This may partially explain why resolution degradation remains a popular (and successful) method for

foveation. Artefacts or perceivable foveation was much more prominent across the ambient method trials, but even within the tested sampling levels we found on which relatively imperceptible and provided substantial computational benefit. Our metric indicates that our ray-casting method is relatively undetectable at lower step rates (but not the lowest). These results may be the first paces towards motivating the use of real-time ray casting content for virtual reality. We expect the computational gains to be even more substantial once we are able to integrate multiple methods together.

We recognize a few limitations of our study. Firstly, we would like to conduct a larger exploration of the parameter space for our rendering methods to make more accurate inferences about the rate of change in terms of detectability. Additionally, we do not explore any temporal aspects of our methods and the detectability any temporal-specific aspects that may be introduced. We realize that temporal evaluation is critical to fully validate foveated methods, requiring accurate, fast, and reliable eye tracking.

Chapter 5

Perception and Hardware

5.1 Dual-Sensor Filtering for Robust HMD Tracking

In 2014, I published a short paper at the 20th ACM Symposium on VRST as primary author in collaboration with other researchers from Disney Research. Our paper was on a novel and inexpensive positional tracking method for head-mounted displays. At the time of submission, available commercial head-mounted displays did not come with positional tracking solutions. Using only the on-board sensors and inexpensive hardware additions we developed a novel system that provided accurate positional tracking in room-sized environments.

We chose an optical tracking system due to how easily accessible the supporting hardware is for ordinary consumers. Our method uses color blob tracking with a wide acceptable color threshold to filter the image space and select candidate regions. We then evaluate a fiducial marker tracking algorithm against the reduced image space and select a candidate. By combining these methods, we overcome the poor performance of color blob tracking under imperfect lighting conditions and the slow processing speeds for traditional fiducial marker tracking (see Figure 5-1). The only additional supporting hardware necessary for this is a printed marker in a vibrant color, which is attached securely to the top of the head-mounted display unit, and a standard high-definition webcam mounted overhead, looking down at the tracked space. Our method supports multiple users and provides both position and orientation data for most standard orientations of the head. We use the positional data as-is. The optical orientation data is Kalman filtered with the IMU's orientation data. By filtering both data streams together, we overcome the optical data's high-frequency but low persistence errors and the IMU's low-frequency but high-persistence errors.

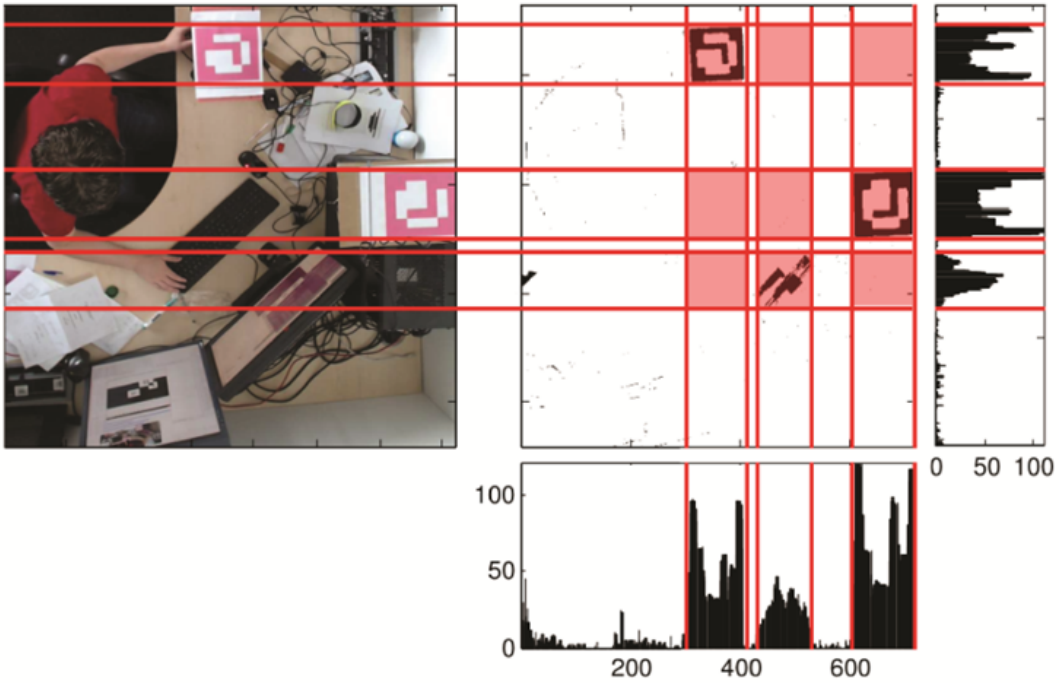

Figure 5-1: The color blob tracking and fiducial marker hybrid approach. We reduce the image search space via a color blob tracking algorithm and then conduct our fiducial marker tracking on the candidate areas. This was further refined to only include areas which were temporally consistent. The robustness of the fiducial marker tracker also meant that we could operate with a larger color threshold and mitigate the effect of poor lighting conditions.

Additionally, our system was networked, allowing users to freely move within the tracked space. A primary machine, acting as a server, carried out optical tracking and transmitted tracking data wirelessly to laptops, which acted as clients, that could be strapped to the users' backs. Each client would then filter the orientation data streams and update the visual's in the head-mounted display accordingly (See Figure 5-2).

Figure 5-2: Our system running across multiple machines. The laptop on the left (server) conducts the image-based tracking. The position and orientation data is wirelessly transmitted to the laptop on the right (client). This data is then Kalman filtered against the HMD's tracking data, which is connected to the client machine.

5.2 Layered Video Format

Also in 2014, I collaborated with researchers at Disney Research on another short-paper for CVMP. The paper presented a novel codec for 3D layered video (See Figure 5-3). The Layered 3D Video format provides depth through parallax. Dynamic content is encoded into a single container, along with additional data for the parallax effect such as the ideal distance between layers. As part of the work, I developed a desktop based viewer that provided parallax using an inexpensive commercial head tracker. This provides fishtank-like virtual/augmented reality most ideally suited for cartoon-like content. I also contributed to the paper writing process and additional supporting code.

Figure 5-3: The layered 3D video format. Each layer can be animated in order to present dynamic content. The illusion of depth is provided via parallax, with layer distances specified in the video container.

5.3 Multi-Monitor Virtual Environment

A major contribution of this work is the development of a commercially viable Virtual Reality Environment (VRE) with perceptual rendering enhancements that take advantage of the system hardware. Due to the low latency and high frame rate requirements to fully explore the variety of perceptual phenomena that have been described so far, our initial approach was to construct a VRE that would take existing traditional hardware that met these requirements and bring them together in an innovative way. It is already common to see gaming enthusiasts construct three to four panel surround systems, and in the financial sector monitor arrays that exceed six monitors are regularly employed. We decided to construct a fully encompassing CAVE composed entirely of low latency, high frame rate, stereoscopic monitors. As far as we are aware, this would have been the first academic endeavor to build such a system, which may have originated a host of novel contributions to the field.

This system, however, has a number of disadvantages. One of our concerns even before assembling the system was how much the monitor bezels would impact the experience. Although previous studies concluded that bezels either don't affect task performance [Bi et al., 2010, De Almeida et al., 2012, McNamara et al., 2011] or sometimes even increase task performance [Grüninger and Krüger, 2013], users reported

Figure 5-4: Image of the prototype multi-monitor VRE. The system used three Asus VG278HE 27" monitors (central portrait and far edges) and four BenQ XL2720Z 27" monitors (flanking the central portrait monitor). The system is running Elite Dangerous at 8760x2160 resolution.

a subjectively worse experience [Hennecke et al., 2012, Bi et al., 2010]. Additionally, bezels are known to have an impact on stereoscopic quality [Grüninger and Krüger, 2013]. Possible solutions to obfuscating bezels are a 'french window' approach which requires tracking [De Almeida et al., 2012] or by projecting obfuscated content onto the bezels [Ebert et al., 2010]. Additionally, working with flat rectangular panels meant that constructing an environment with some curvature would incur gaps between the panels in addition to the bezels. After construction, our own subjective experience with the system suggested that the combination of gaps and bezels would be too pronounced, especially for stereoscopic rendering, although in hindsight a 'french window' approach may have alleviated the effect somewhat.

Our primary concern, however, was the cost for a fully immersive system and the limited benefits the underlying hardware would yield. This was due in equal parts to the high specifications of the monitors and the sturdiness of the supporting structure. Our prototype used a relatively cheap to construct frame, but as a consequence was very susceptible to bumps and stress at the joints. The material engineering effort required coupled with both the limited benefit and lack of flexibility of this system led to its eventual abandonment and dismantling in favor of an alternative.

5.4 Multi-Projector Virtual Environment

Our alternative, therefore, was to develop a multiple projector based system with pan-tilt projection redirection that can achieve saccadic speeds in the style of [Iwai et al., 2015] and [Okumura et al., 2011]. In comparison to the multi-monitor approach, a multi-projector solution can be substantially cheaper while allowing finer control over foveal display quality. With just two 1080p projectors we should be able to simulate 8K or higher displays at a fraction of the cost that a single projector with equivalent capability would market for. Additionally, given our control over the actual projection cone, the foveal display could be condensed such that the pixel density approached MSA under typical viewing conditions.

Figure 5-5: Full shot of our prototype pan-tilt projection steering system. The system is currently only capable of redirecting on one axis.

In order to redirect projection, we may use two first-surface mirrors mounted on separate actuators (which requires a narrower foveal projection frustrum) or a single mirror actuated on both pan and tilt (requires additional frame engineering). We currently have not decided on either approach, which will require further evaluation with our existing hardware. The motors and respective controlling boards are able to achieve saccadic speeds with minimal command latency (sub millisecond according to manufacturers). We have developed a prototypical system (see Figures 5-5 and 5-6) that is currently able to redirect foveal content on a single axis.

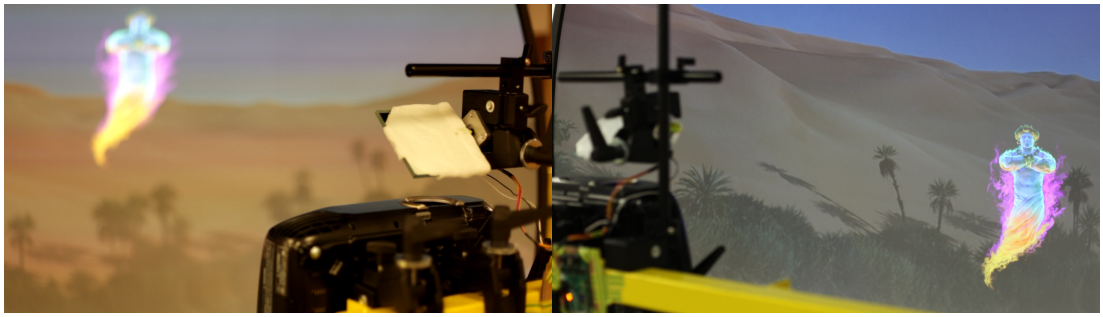

Figure 5-6: Stills of the projection steering system in action, with focus on the mechanical parts (left) and the redirected projection (right).

However, such a system is not without its cost, especially in the immersive VREs we are targeting. The distortions from both the pan-tilt redirection and the display surface must be taken into consideration, so significant calibration is required [Nakamura and Hiraike, 2002, Ashdown and Sato, 2005, Mitsugami et al., 2005]. This distortion correction will reduce the effective working resolution at higher incident angles (away from 45°) between the projector and the mirror. Additionally, blending between the foveal and peripheral projections will be a significant concern especially due to the difference in luminance between projections (due to hardware differences and projection areas), transparency concerns (see Figure 5-7), and the need for accurate registration [Raskar et al., 1999].

The development of this system will allow me to provide in-situ examples of the effectiveness of my perceptually lossless research where there are large benefits in computational savings. Once the test-bed is constructed, focus can shift first and foremost to the development of new rendering models, some of which may exploit the inherent properties of the hardware. Building on prior work, I expect that the construction of a full system is definitely achievable within the time left to complete my doctorate.

Figure 5-7: Close up view of the quality difference between projections. The foveal projector, displaying the genie, operates at a much higher pixel density than the peripheral projection, displaying the desert background. Although visible blending and transparency issues and transparency could be somewhat alleviated by a mask in the peripheral projection, there will always be a need for blending on window edges as the content refresh rate of the projector (60 Hz) cannot match the speed of the motors and may cause visible artifacts after a saccade.

5.5 Eye Tracking

Given that many aspects of the work rely on third-party components or systems, it may seem strange to dedicate engineering effort to the construction of an eye tracking system. The reason why is quite pragmatic: academic level eye tracking solutions are prohibitively expensive and inflexible, often tied to a host machine and monitor to which the system is optimized for; commercial level tracking is, in contrast, much cheaper but does not provide the level of quality required to achieve perceptually lossless results. An extensive search of eye tracking solutions led to the conclusion that a novel and flexible system was required to fit the previously discussed use-cases. An added benefit is that a novel, in-house system allows control over all critical aspects of the tracking pipeline and system design.

Figure 5-8: Stills of an eye under off-axis NIR illumination (850 nm) from the prototype head-mount (no hot mirror). Under visible light the iris would be dark brown. Under NIR light, the iris is much lighter than the dark pupil. Corneal reflection of the two fixed IRED light sources are visible in the pupil. Note that shadowing from the eyelids and eyelashes can still cause interference.

One of the critical aspects of this solution is that it must be as unobtrusive as possible to use in order to truly evaluate the effectiveness of novel perceptual rendering methods. Additionally, the nature and purpose of VREs meant that a remote eye tracking system would be difficult to adopt unless we could accurately account for every position and orientation a user could make within the environment. A head-mounted solution seemed like the obvious choice, but generally head-mounted solutions sacrifice some of the user's field of view for better camera placement.

In order to overcome this issue, we opted to use a hot mirror based solution to reduce visual interference as much as possible. Hot mirrors have high reflectance rates for NIR wavelengths while simultaneously providing high transmission rates for all other light sources (although typically specialized for one band, such as visible light). Hot mirror eye tracking solutions have been employed in academic [Boening et al., 2006, Cho et al., 2009] and commercial[1] settings prior. The major advantage of these systems is that they allow for ideal capture angles with minimal interference, as the camera can be mounted outside the user's field of vision.

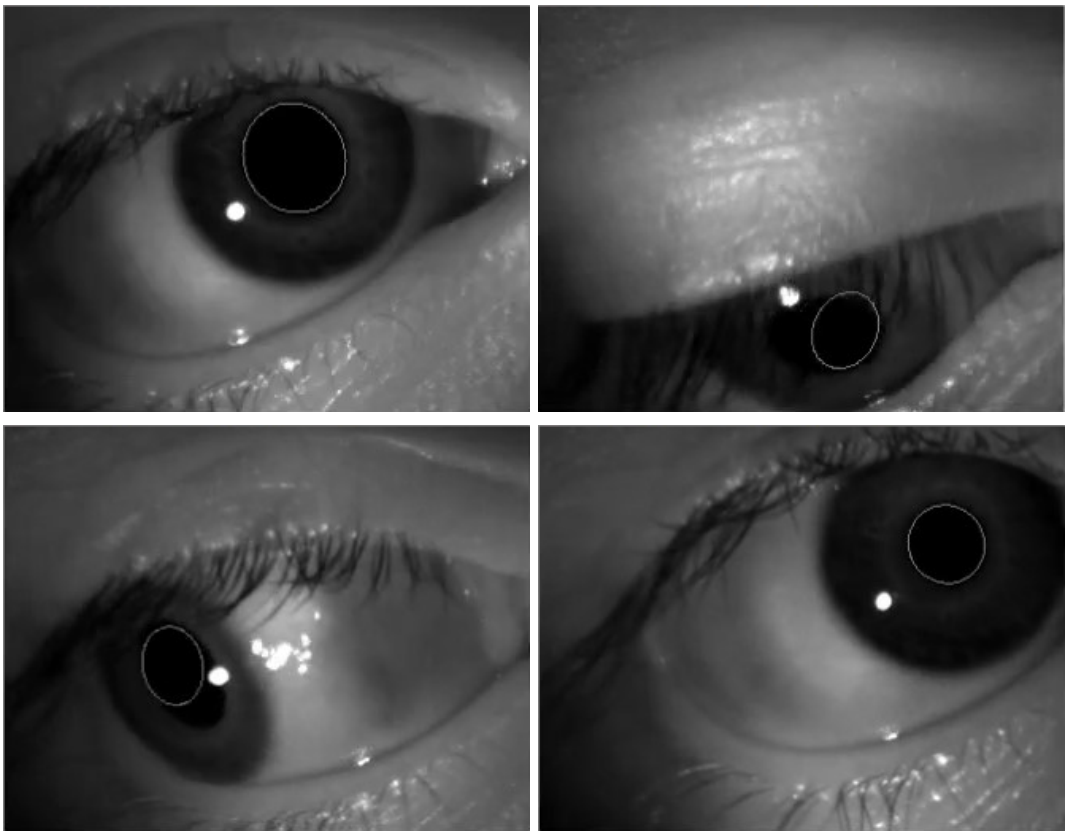

Figure 5-9: Several stills showing Starburst ellipse fitting on visible light footage from the reimplementation of the algorithm in C++. The ellipse (in grey) fits tightly to the pupillary edge when successful. The algorithm fails when there is heavier interference from the eyelashes.

On the software side, the Starburst algorithm discussed in Section 5.5 was reimplemented in C++. Figure 5-9 shows the algorithm in action with visible light feed of the left eye. As expected, there are issues with eyelashes occluding the pupil and tracking

[1]See commercial implementation of EyeSeeCam (http://www.interacoustics.com/eyeseecam) and EyeGuide (http://eye.guide/hardware) for examples.

failure at high eccentricities or significantly dark footage. This is somewhat alleviated by using NIR footage as the illumination can be increased without bothering the users and the cornea lightens making the pupil more distinguishable from its surroundings (see Figure 5-8).

Figure 5-10: Virtual model (left) and physical model (right) of our prototype head-mounted eye tracking frame, developed with help from resident Disney artists. Pictured also is an IDS UI-3370CP NIR machine vision camera, which is the model being used for our solution. It is equipped with a 12mm FL lens.

In collaboration with Disney Research artists, we also spent time designing the frame supporting the eye tracking hardware, as can be seen in Figure 5-10. The eye tracking portion of this work has scope outside the project and, potentially, applicability within other business areas of Disney. As such, the frame should be functional, ergonomic, and aesthetically pleasing to justify commercial adoption. At later stages of development, this will require further investigation into human factors engineering, interaction, and ergonomic literature.

Chapter 6

Conclusion

6.1 Consolidation

The ultimate goal of the project would be to develop a perceptually aware virtual environment, based on projection redirection according to user gaze. The head-mounted eye tracker would allow the user significant mobility within the volume. The addition of a head-tracking device would be comparatively simple. Figure 6-1 shows a system diagram that clarifies the relevant connections in the full Perceptually Augmented Virtual Environment (PAVE) system planned for this project. The original goal of the project may seem massive given the time-frame of a Masters degree, but this project was initially intended for a Engineering Doctorate. Due to personal reasons, this had to be abandoned to pursue a Masters degree in its stead.

As mentioned, projection redirection for foveated rendering on a curved projection surface requires significant amount of distortion correction and significant amounts of projection registration (which could be pre-calibrated since the surface is not dynamic). The driving force for the mirrors must be fast and robust enough to undergo near-saccadic speeds. Additionally, communication latency between the tracking devices, the host machine, and the projection system must be minimal. Simultaneously, work is conducted on implementing existing and developing new foveated rendering methods under a unified rendering framework. Once the system is complete, we are able to test foveated rendering methods extensively in a commercially viable framework.

One of the key aspects of the engineering effort behind the project is cost reduction. This is evident by many of the design choices made for the VRE, particularly the multiple projector system. Instead of relying on very high quality projectors (e.g. 4K 120Hz projectors, which can breeze past thousands of dollars) the system would simulate that quality level through the use of two lower quality (and consequently cheaper)

projectors and some clever, responsive positioning. Additionally, by actuating mirrors rather than the projector itself, cost is reduced further at the expense of increased distortion correction complexity.

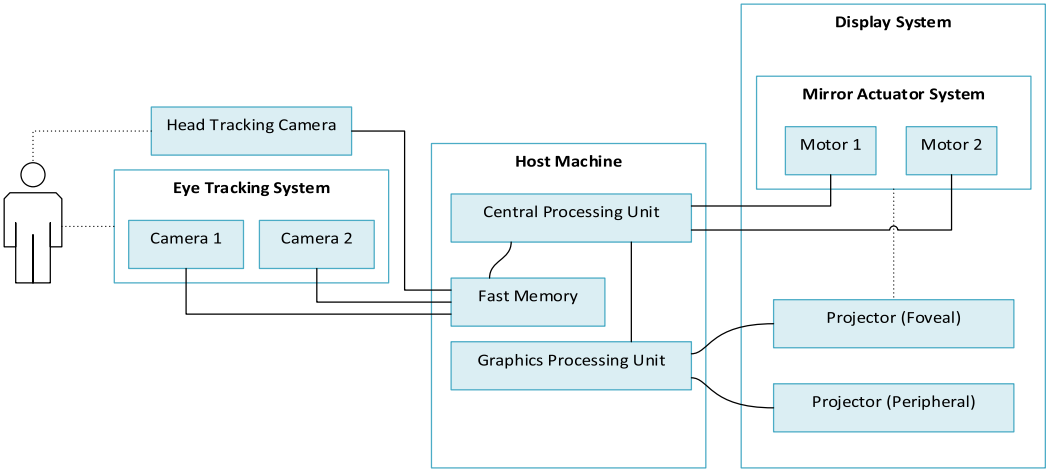

Figure 6-1: System diagram clarifying communication links between elements in the system. Solid lines represent digital communication pathways within the system, while dotted lines represent functional connections (e.g. the mirror actuator system redirects the foveal projection).

Some of these design choices were also made to be easily also portable to existing VR hardware. The hot-mirror eye-tracking design is portable (perhaps ideal) for head-mounted displays. The controlled lighting environment and no camera size restrictions (within reason) means there is no significant size-quality trade-off. Of course, other limitations (such as weight) are still present, but they would not be significantly different from the limitations with a standalone head-mounted eye tracker.

6.2 Future Work

There is a push towards greater advancements in real-time rendering, particularly due to the demands of virtual reality. As the field begins to closely integrate with human sensory limits, it seems logical not only to account for but also exploit perceptual systems. This body work serves as an initial foray into a closer and more efficient integration of human perception and simulated reality. Further advances on the topic will lead to more energy efficient, economic accessible, and more refined focus for real-time rendering in the entertainment industry.

Future work includes the development and testing of additional foveated rendering

methods which explore the perceptual space fully. This work contains some preliminary steps in the right direction, but further studies require more representative user bases and robust systems. Machine learning for gaze prediction provides a sufficiently deep and interesting challenge to form its own research topic. Similarly, perceptually aware hardware engineering is another logical and valuable segue from this work.

An immediate follow up to this work is the completion of the PAVE including the development, validation, and integration of the gaze-contingent rendering methods introduced. A fully constructed system will provide many benefits: Firstly, it provides a commercially viable framework that can refined to meet consumer cost demands; Secondly, it provides a framework to test current and future methods effectively; Finally, it can provide better academic insight about the human perceptual system through comprehensive user tracking.

6.3 Summary

This thesis outlines the value of considering human perception for real-time rendering applications, particularly in the current commercial and academic climate. It also contains a substantial review of optical physiology, psychophysics, virtual reality, and perceptually based real-time rendering literature that is immediately relevant to the work at hand. Additionally, it outlines the initial steps and considerations for the construction of a perceptually aware virtual reality environment that would fully exploit, and clearly benefit from, gaze-contingent rendering. The work conducted to date is presented and its tangible output, both academically and industrially, is discussed. The research questions posed in Section 1.3 are answered incrementally through the various projects undertaken throughout the research.

In the same way that dedicated graphics processing units were once a novelty, passive user-tracking may eventually become ubiquitous once prices lower and quality increases. As we approach the limits of rendering hardware and algorithmic hacks and wait for the next leap forward, it may be time to look at advances in other fields to assist our goals in real-time rendering. Through user-tracking, we may push a new paradigm for what it means to be smart about rendering. Ideally, of course, we wouldn't have to rely on clever hacks or work arounds; our aim is to replicate how things work in nature as faithfully as possible. However if there is any claim that the current ubiquity of rasterization supports it is that, in reality, practice is rarely as ideal as theory but demand always asks for more than what we thought was possible in practice.

Hopefully, the work in this thesis motivates others to pursue a deeper study of perception in computer graphics, especially with the advent of commercial virtual reality.

Many in the field have shown how important human factors are in the field, and its commercialization has served to shed even more light on its importance. Instead of just accounting for human perception and sensibility, this (and all the existing work in the field) shows the importance and benefit of *using* human perception to our advantage.

Glossary

AR Augmented Reality. 73

CAVE Cave Automatic/Computer Assisted Virtual Environment. 24, 73

CDE Centre for Digital Entertainment. 73

CFF Critical Flicker Frequency. 15–17, 73

CGI Computer-Generated Imagery. 73

CMF Cortical Magnification Factor. 14, 50, 57, 73

CovSAL Erdem and Erdem. 54, 73

CPU Central Processing Unit. 73

CRT Cathode Ray Tube. 73

CSF Contrast Sensitivity Function. 18, 33, 50, 73

CVMP Conference on Visual Media Production. 6, 62, 73

DEN Digital Economy Network. 73

EPSRC Engineering and Physical Sciences Research Council. 73

FA-SSIM Foveation-based content Adaptive SSIM. 38, 73

F-MSE Foveated Mean-Squared Error. 73

FPS Frame per Second. 73

FSNR Foveated Signal-to-Noise Ratio. 38, 73

F-SSIM Foveated Structural Similarity Index. 73

F-UQI Foveated Universal Image Quality Index. 73

FWQI Foveated Wavelet Quality Index. 37, 73

GBVS Graph-Based Visual Saliency. 53, 54, 73

GPU Graphics Processing Unit. 29, 30, 73

HDR-VDP2 HDR Visual Difference Predictor. 37, 47, 50, 52, 57, 58, 73

HMD Head-Mounted Display. 4, 24, 30, 44, 62, 73

HVS Human Visual System. 73

I3D Interactive 3D Graphics and Games. 73

IMU Inertial Measurement Unit. 73

IPD Interpupillary Distance. 73

IRED Infrared Emitting Diode. 73

ITTI Itti, Koch, and Neibur. 54, 73

JND just noticeable difference. 73

LED Light Emitting Diode. 73

LOD level-of-detail. 30, 73

MSA Minimum Separable Acuity. 18, 65, 73

MSE Mean-Squared Error. 73

NIR Near-Infrared (Illumination). 22, 23, 38, 69, 70, 73

PAVE Perceptually Augmented Virtual Environment. 71, 73

PSNR Peak Signal-to-Noise Ratio. 38, 73

SSAO Screen-Space Ambient Occlusion. 47, 52, 53, 57–59, 73

SSIM Structural Similarity Index. 37, 38, 73

unreal Unreal Engine 4. 73

UQI Universal Image Quality Index. 73

VR Virtual Reality. 4, 6, 7, 24, 30, 38, 39, 72, 73

VRE Virtual Reality Environment. 7, 24, 39, 63, 66, 68, 71, 73

VRST Virtual Reality Software and Technology. 5, 60, 73

Bibliography

[Abrams et al., 1989] Abrams, R. A., Meyer, D. E., and Kornblum, S. (1989). Speed and accuracy of saccadic eye movements: characteristics of impulse variability in the oculomotor system. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):529.

[Alpern and Spencer, 1953] Alpern, M. and Spencer, R. W. (1953). Variation of critical flicker frequency in the nasal visual field: Relation to variation in size of the entrance pupil and to stray light within the eye. *AMA archives of ophthalmology*, 50(1):50–63.

[and others, 1962] and others (1962). Sensorama simulator. US Patent 3,050,870.

[Anderson et al., 1991] Anderson, S., Mullen, K., and Hess, R. (1991). Human peripheral spatial resolution for achromatic and chromatic stimuli: limits imposed by optical and retinal factors. *The Journal of Physiology*, 442(1):47–64.

[Ashdown and Sato, 2005] Ashdown, M. and Sato, Y. (2005). Steerable projector calibration. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 98–98. IEEE.

[Bahill et al., 1975] Bahill, A. T., Adler, D., and Stark, L. (1975). Most naturally occurring human saccades have magnitudes of 15 degrees or less. *Investigative ophthalmology*, 14(6):468.

[Bahill and Stark, 1979] Bahill, A. T. and Stark, L. (1979). The trajectories of saccadic eye movements. *Scientific American*, 240(1):108–117.

[Bailey et al., 2009] Bailey, R., McNamara, A., Sudarsanam, N., and Grimm, C. (2009). Subtle gaze direction. *ACM Transactions on Graphics (TOG)*, 28(4):100.

[Baloh et al., 1975] Baloh, R. W., Konrad, H. R., Sills, A. W., and Honrubia, V. (1975). The saccade velocity test. *Neurology*, 25(11):1071–1071.

[Baudisch et al., 2003] Baudisch, P., DeCarlo, D., Duchowski, A. T., and Geisler, W. S. (2003). Focusing on the essential: considering attention in display design. *Communications of the ACM*, 46(3):60–66.

[Bavoil and Sainz, 2008] Bavoil, L. and Sainz, M. (2008). Screen space ambient occlusion. *NVIDIA developer information: http://developers. nvidia. com*, 6.

[Bi et al., 2010] Bi, X., Bae, S.-H., and Balakrishnan, R. (2010). Effects of interior bezels of tiled-monitor large displays on visual search, tunnel steering, and target selection. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 65–74. ACM.

[Boening et al., 2006] Boening, G., Bartl, K., Dera, T., Bardins, S., Schneider, E., and Brandt, T. (2006). Mobile eye tracking as a basis for real-time control of a gaze driven head-mounted video camera. In *Proceedings of the 2006 symposium on Eye tracking research & applications*, pages 56–56. ACM.

[Boghen et al., 1974] Boghen, D., Troost, B., Daroff, R., Dell'Osso, L., and Birkett, J. (1974). Velocity characteristics of normal human saccades. *Investigative Ophthalmology*, 13(8):619.

[Borji and Itti, 2014] Borji, A. and Itti, L. (2014). Defending yarbus: Eye movements reveal observers' task. *Journal of vision*, 14(3):29.

[Brown et al., 2005] Brown, M., Majumder, A., and Yang, R. (2005). Camera-based calibration techniques for seamless multiprojector displays. *Visualization and Computer Graphics, IEEE Transactions on*, 11(2):193–206.

[Campbell and Westheimer, 1960] Campbell, F. and Westheimer, G. (1960). Dynamics of accommodation responses of the human eye. *The Journal of physiology*, 151(2):285–295.

[Cater et al., 2002] Cater, K., Chalmers, A., and Ledda, P. (2002). Selective quality rendering by exploiting human inattentional blindness: looking but not seeing. In *Proceedings of the ACM symposium on Virtual reality software and technology*, pages 17–24. ACM.

[Cave and Bichot, 1999] Cave, K. R. and Bichot, N. P. (1999). Visuospatial attention: Beyond a spotlight model. *Psychonomic Bulletin & Review*, 6(2):204–223.

[Chen et al., 2000] Chen, Y., Clark, D. W., Finkelstein, A., Housel, T. C., and Li, K. (2000). Automatic alignment of high-resolution multi-projector display using an

un-calibrated camera. In *Proceedings of the conference on Visualization'00*, pages 125–130. IEEE Computer Society Press.

[Cho et al., 2009] Cho, C. W., Lee, J. W., Lee, E. C., and Park, K. R. (2009). Robust gaze-tracking method by using frontal-viewing and eye-tracking cameras. *Optical Engineering*, 48(12):127202–127202.

[Cowey and Rolls, 1974] Cowey, A. and Rolls, E. (1974). Human cortical magnification factor and its relation to visual acuity. *Experimental Brain Research*, 21(5):447–454.

[Creed and Ruch, 1932] Creed, R. and Ruch, T. C. (1932). Regional variations in sensitivity to flicker. *The Journal of physiology*, 74(4):407–423.

[Cruz-Neira et al., 1993] Cruz-Neira, C., Sandin, D. J., and DeFanti, T. A. (1993). Surround-screen projection-based virtual reality: the design and implementation of the cave. In *Proceedings of the 20th annual conference on Computer graphics and interactive techniques*, pages 135–142. ACM.

[Cruz-Neira et al., 1992] Cruz-Neira, C., Sandin, D. J., DeFanti, T. A., Kenyon, R. V., and Hart, J. C. (1992). The cave: audio visual experience automatic virtual environment. *Communications of the ACM*, 35(6):64–72.

[Curcio and Allen, 1990] Curcio, C. A. and Allen, K. A. (1990). Topography of ganglion cells in human retina. *Journal of Comparative Neurology*, 300(1):5–25.

[Curcio et al., 1991] Curcio, C. A., Allen, K. A., Sloan, K. R., Lerea, C. L., Hurley, J. B., Klock, I. B., and Milam, A. H. (1991). Distribution and morphology of human cone photoreceptors stained with anti-blue opsin. *Journal of Comparative Neurology*, 312(4):610–624.

[Curry et al., 2003] Curry, D. G., Martinsen, G. L., and Hopper, D. G. (2003). Capability of the human visual system. In *AeroSense 2003*, pages 58–69. International Society for Optics and Photonics.

[De Almeida et al., 2012] De Almeida, R. A., Pillias, C., Pietriga, E., and Cubaud, P. (2012). Looking behind bezels: French windows for wall displays. In *Proceedings of the International Working Conference on Advanced Visual Interfaces*, pages 124–131. ACM.

[Dougherty et al., 2003] Dougherty, R. F., Koch, V. M., Brewer, A. A., Fischer, B., Modersitzki, J., and Wandell, B. A. (2003). Visual field representations and locations of visual areas v1/2/3 in human visual cortex. *Journal of Vision*, 3(10):1.

[Duchowski and Coltekin, 2007] Duchowski, A. T. and Coltekin, A. (2007). Foveated gaze-contingent displays for peripheral lod management, 3d visualization, and stereo imaging. *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMCCAP)*, 3(4):6.

[Ebert et al., 2010] Ebert, A., Thelen, S., Olech, P.-S., Meyer, J., and Hagen, H. (2010). Tiled++: An enhanced tiled hi-res display wall. *Visualization and Computer Graphics, IEEE Transactions on*, 16(1):120–132.

[Ebisawa, 1998] Ebisawa, Y. (1998). Improved video-based eye-gaze detection method. *Instrumentation and Measurement, IEEE Transactions on*, 47(4):948–955.

[Erdem and Erdem, 2013] Erdem, E. and Erdem, A. (2013). Visual saliency estimation by nonlinearly integrating features using region covariances. *Journal of vision*, 13(4):11.

[Febretti et al., 2013] Febretti, A., Nishimoto, A., Thigpen, T., Talandis, J., Long, L., Pirtle, J., Peterka, T., Verlo, A., Brown, M., Plepys, D., et al. (2013). Cave2: a hybrid reality environment for immersive simulation and information analysis. In *IS&T/SPIE Electronic Imaging*, pages 864903–864903. International Society for Optics and Photonics.

[Ferry, 1892] Ferry, E. S. (1892). Persistence of vision. *American Journal of Science*, (261):192–207.

[Fischler and Bolles, 1981] Fischler, M. A. and Bolles, R. C. (1981). Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395.

[Fukuda, 1979] Fukuda, T. (1979). Relation between flicker fusion threshold and retinal positions. *Perceptual and motor skills*, 49(1):3–17.

[Gibson et al., 1959] Gibson, E. J., Gibson, J. J., Smith, O. W., and Flock, H. (1959). Motion parallax as a determinant of perceived depth. *Journal of experimental psychology*, 58(1):40.

[Granit and Harper, 1930] Granit, R. and Harper, P. (1930). Comparative studies on the peripheral and central retina. *American Journal of Physiology–Legacy Content*, 95(1):211–228.

[Grüninger and Krüger, 2013] Grüninger, J. and Krüger, J. (2013). The impact of display bezels on stereoscopic vision for tiled displays. In *Proceedings of the 19th ACM Symposium on Virtual Reality Software and Technology*, pages 241–250. ACM.

[Guenter et al., 2012] Guenter, B., Finch, M., Drucker, S., Tan, D., and Snyder, J. (2012). Foveated 3d graphics. *ACM Transactions on Graphics (TOG)*, 31(6):164.

[Hagstrom et al., 1998] Hagstrom, S. A., Neitz, J., and Neitz, M. (1998). Variations in cone populations for red–green color vision examined by analysis of mrna. *NeuroReport*, 9(9):1963–1967.

[Harel et al., 2006] Harel, J., Koch, C., and Perona, P. (2006). Graph-based visual saliency. In *Advances in neural information processing systems*, pages 545–552.

[Hecht and Verrijp, 1933] Hecht, S. and Verrijp, C. D. (1933). Intermittent stimulation by light iii. the relation between intensity and critical fusion frequency for different retinal locations. *The Journal of general physiology*, 17(2):251–268.

[Hennecke et al., 2012] Hennecke, F., Matzke, W., and Butz, A. (2012). How screen transitions influence touch and pointer interaction across angled display arrangements. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 209–212. ACM.

[Henriksson et al., 1980] Henriksson, N., Pyykko, I., Schalen, L., and Wennmo, C. (1980). Velocity patterns of rapid eye movements. *Acta oto-laryngologica*, 89(3-6):504–512.

[Hofer et al., 2005] Hofer, H., Carroll, J., Neitz, J., Neitz, M., and Williams, D. R. (2005). Organization of the human trichromatic cone mosaic. *The Journal of Neuroscience*, 25(42):9669–9679.

[Holmes et al., 1977] Holmes, D. L., Cohen, K. M., Haith, M. M., and Morrison, F. J. (1977). Peripheral visual processing. *Perception & Psychophysics*, 22(6):571–577.

[Hoppe, 1996] Hoppe, H. (1996). Progressive meshes. In *Proceedings of the 23rd annual conference on Computer graphics and interactive techniques*, pages 99–108. ACM.

[Hutchinson et al., 1989] Hutchinson, T. E., White Jr, K. P., Martin, W. N., Reichert, K. C., Frey, L., et al. (1989). Human-computer interaction using eye-gaze input. *Systems, Man and Cybernetics, IEEE Transactions on*, 19(6):1527–1534.

[Hylkema, 1942] Hylkema, B. (1942). Fusion frequency with intermittent light under various circumstances. *Acta Ophthalmologica*, 20(2):159–180.

[Itti, 2004] Itti, L. (2004). Automatic foveation for video compression using a neurobiological model of visual attention. *Image Processing, IEEE Transactions on*, 13(10):1304–1318.

[Itti et al., 1998] Itti, L., Koch, C., and Niebur, E. (1998). A model of saliency-based visual attention for rapid scene analysis. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (11):1254–1259.

[Iwai et al., 2015] Iwai, D., Kodama, K., and Sato, K. (2015). Reducing motion blur artifact of foveal projection for a dynamic focus-plus-context display. *Circuits and Systems for Video Technology, IEEE Transactions on*, 25(4):547–556.

[Jaeger, 2009] Jaeger, C. (2009). Eye safety of ireds used in lamp applications. *Application note, OSRAM Opto Semiconductors GmbH*.

[Jones et al., 2014] Jones, B., Sodhi, R., Murdock, M., Mehra, R., Benko, H., Wilson, A., Ofek, E., MacIntyre, B., Raghuvanshi, N., and Shapira, L. (2014). Roomalive: Magical experiences enabled by scalable, adaptive projector-camera units. In *Proceedings of the 27th annual ACM symposium on User interface software and technology*, pages 637–644. ACM.

[Kolasinski, 1995] Kolasinski, E. M. (1995). Simulator sickness in virtual environments. Technical report, DTIC Document.

[Kuroki et al., 2007] Kuroki, Y., Nishi, T., Kobayashi, S., Oyaizu, H., and Yoshimura, S. (2007). A psychophysical study of improvements in motion-image quality by using high frame rates. *Journal of the Society for Information Display*, 15(1):61–68.

[Lee et al., 1999] Lee, S., Bovik, A. C., and Kim, Y. Y. (1999). Low delay foveated visual communications over wireless channels. In *Image Processing, 1999. ICIP 99. Proceedings. 1999 International Conference on*, volume 3, pages 90–94. IEEE.

[Lee et al., 2002] Lee, S., Pattichis, M. S., and Bovik, A. C. (2002). Foveated video quality assessment. *Multimedia, IEEE Transactions on*, 4(1):129–132.

[Legge and Foley, 1980] Legge, G. E. and Foley, J. M. (1980). Contrast masking in human vision. *JOSA*, 70(12):1458–1471.

[Legge and Kersten, 1987] Legge, G. E. and Kersten, D. (1987). Contrast discrimination in peripheral vision. *JOSA A*, 4(8):1594–1598.

[Levoy and Whitaker, 1990] Levoy, M. and Whitaker, R. (1990). Gaze-directed volume rendering. *ACM SIGGRAPH Computer Graphics*, 24(2):217–223.

[Li et al., 2005] Li, D., Winfield, D., and Parkhurst, D. J. (2005). Starburst: A hybrid algorithm for video-based eye tracking combining feature-based and model-based

approaches. In *Computer Vision and Pattern Recognition-Workshops, 2005. CVPR Workshops. IEEE Computer Society Conference on*, pages 79–79. IEEE.

[Li et al., 2011] Li, Z., Qin, S., and Itti, L. (2011). Visual attention guided bit allocation in video compression. *Image and Vision Computing*, 29(1):1–14.

[Loftus and Mackworth, 1978] Loftus, G. R. and Mackworth, N. H. (1978). Cognitive determinants of fixation location during picture viewing. *Journal of Experimental Psychology: Human perception and performance*, 4(4):565.

[Lythgoe and Tansley, 1929] Lythgoe, R. and Tansley, K. (1929). The relation of the critical frequency of flicker to the adaptation of the eye. *Proceedings of the Royal Society of London. Series B, Containing Papers of a Biological Character*, pages 60–92.

[Mackworth and Morandi, 1967] Mackworth, N. H. and Morandi, A. J. (1967). The gaze selects informative details within pictures. *Perception & Psychophysics*, 2(11):547–552.

[Mannan et al., 1995] Mannan, S., Ruddock, K., and Wooding, D. (1995). Automatic control of saccadic eye movements made in visual inspection of briefly presented 2-d images. *Spatial vision*, 9(3):363–386.

[Mantiuk et al., 2011] Mantiuk, R., Kim, K. J., Rempel, A. G., and Heidrich, W. (2011). Hdr-vdp-2: a calibrated visual metric for visibility and quality predictions in all luminance conditions. In *ACM Transactions on Graphics (TOG)*, volume 30, page 40. ACM.

[McNamara et al., 2008] McNamara, A., Bailey, R., and Grimm, C. (2008). Improving search task performance using subtle gaze direction. In *Proceedings of the 5th symposium on Applied perception in graphics and visualization*, pages 51–56. ACM.

[McNamara et al., 2011] McNamara, A., Parke, F., and Sanford, M. (2011). Exploring the effect of tiling on large displays. In *SIGGRAPH Asia 2011 Posters*, page 57. ACM.

[Mitsugami et al., 2005] Mitsugami, I., Ukita, N., and Kidode, M. (2005). Fixed-center pan-tilt projector and its calibration methods. *MVA*, 5:492–497.

[Murphy et al., 2009] Murphy, H. A., Duchowski, A. T., and Tyrrell, R. A. (2009). Hybrid image/model-based gaze-contingent rendering. *ACM Transactions on Applied Perception (TAP)*, 5(4):22.

[Myers et al., 1991] Myers, G., Sherman, K. R., Stark, L., et al. (1991). Eye monitor: microcomputer-based instrument uses an internal mode to track the eye. *Computer*, 24(3):14–21.

[Nakamura and Hiraike, 2002] Nakamura, N. and Hiraike, R. (2002). Active projector: Image correction for moving image over uneven screens. In *Companion of the 15th Annual ACM Symposium on User Interface Software and Technology*, pages 1–2.

[Ohno et al., 2002] Ohno, T., Mukawa, N., and Yoshikawa, A. (2002). Freegaze: a gaze tracking system for everyday gaze interaction. In *Proceedings of the 2002 symposium on Eye tracking research & applications*, pages 125–132. ACM.

[Ohshima et al., 1996] Ohshima, T., Yamamoto, H., and Tamura, H. (1996). Gaze-directed adaptive rendering for interacting with virtual space. In *Virtual Reality Annual International Symposium, 1996., Proceedings of the IEEE 1996*, pages 103–110. IEEE.

[Okumura et al., 2011] Okumura, K., Oku, H., and Ishikawa, M. (2011). High-speed gaze controller for millisecond-order pan/tilt camera. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 6186–6191. IEEE.

[Okumura et al., 2013] Okumura, K., Oku, H., and Ishikawa, M. (2013). Acitve projection ar using high-speed optical axis control and appearance estimation algorithm. In *Multimedia and Expo (ICME), 2013 IEEE International Conference on*, pages 1–6. IEEE.

[OpenStax College, 2013] OpenStax College (2013). *Anatomy & Physiology*.

[Osterberg, 1935] Osterberg, G. (1935). *Topography of the layer of rods and cones in the human retina*. Nyt Nordisk Forlag.

[O'Sullivan and Dingliana, 2001] O'Sullivan, C. and Dingliana, J. (2001). Collisions and perception. *ACM Transactions on Graphics (TOG)*, 20(3):151–168.

[Parkhurst and Niebur, 2004] Parkhurst, D. and Niebur, E. (2004). A feasibility test for perceptually adaptive level of detail rendering on desktop systems. In *Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, pages 49–56. ACM.

[Peters and Itti, 2007] Peters, R. and Itti, L. (2007). Congruence between model and human attention reveals unique signatures of critical visual events. In *Advances in neural information processing systems*, pages 1145–1152.

[Pharr and Green, 2004] Pharr, M. and Green, S. (2004). Ambient occlusion. *GPU Gems*, 1:279–292.

[Pirenne, 1967] Pirenne, M. H. (1967). *Vision and the Eye*. Chapman and Hall London.

[Poggel et al., 2006] Poggel, D. A., Treutwein, B., Calmanti, C., and Strasburger, H. (2006). Increasing the temporal g (r) ain: Double-pulse resolution is affected by the size of the attention focus. *Vision Research*, 46(18):2998–3008.

[Policarpo and Oliveira, 2006] Policarpo, F. and Oliveira, M. M. (2006). Relief mapping of non-height-field surface details. In *Proceedings of the 2006 symposium on Interactive 3D graphics and games*, pages 55–62. ACM.

[Polyak, 1941] Polyak, S. L. (1941). The retina.

[Porter, 1902] Porter, T. C. (1902). Contributions to the study of flicker. paper ii. *Proceedings of the Royal Society of London*, 70(459-466):313–329.

[Proffitt and Gilden, 1989] Proffitt, D. R. and Gilden, D. L. (1989). Understanding natural dynamics. *Journal of Experimental Psychology: Human Perception and Performance*, 15(2):384.

[Raskar et al., 1999] Raskar, R., Brown, M. S., Yang, R., Chen, W.-C., Welch, G., Towles, H., Scales, B., and Fuchs, H. (1999). Multi-projector displays using camera-based registration. In *Visualization'99. Proceedings*, pages 161–522. IEEE.

[Raskar et al., 2006] Raskar, R., Van Baar, J., Beardsley, P., Willwacher, T., Rao, S., and Forlines, C. (2006). ilamps: geometrically aware and self-configuring projectors. In *ACM SIGGRAPH 2006 Courses*, page 7. ACM.

[Reinagel and Zador, 1999] Reinagel, P. and Zador, A. M. (1999). Natural scene statistics at the centre of gaze. *Network: Computation in Neural Systems*, 10(4):341–350.

[Reingold and Loschky, 2002] Reingold, E. M. and Loschky, L. C. (2002). Saliency of peripheral targets in gaze-contingent multiresolutional displays. *Behavior Research Methods, Instruments, & Computers*, 34(4):491–499.

[Riddell, 1936] Riddell, L. (1936). The use of the flicker phenomenon in the investigation of the field of vision. *The British journal of ophthalmology*, 20(7):385.

[Rimac-Drlje et al., 2010] Rimac-Drlje, S., Vranješ, M., and Žagar, D. (2010). Foveated mean squared errora novel video quality metric. *Multimedia tools and applications*, 49(3):425–445.

[Roorda and Williams, 1999] Roorda, A. and Williams, D. R. (1999). The arrangement of the three cone classes in the living human eye. *Nature*, 397(6719):520–522.

[Ross, 1936] Ross, R. T. (1936). The fusion frequency in different areas of the visual field: Ii. the regional gradient of fusion frequency. *The Journal of General Psychology*, 15(1):161–170.

[Rovamo and Virsu, 1979] Rovamo, J. and Virsu, V. (1979). An estimation and application of the human cortical magnification factor. *Experimental Brain Research*, 37(3):495–510.

[Ryan et al., 2008] Ryan, W. J., Woodard, D. L., Duchowski, A. T., and Birchfield, S. T. (2008). Adapting starburst for elliptical iris segmentation. In *Biometrics: Theory, Applications and Systems, 2008. BTAS 2008. 2nd IEEE International Conference on*, pages 1–7. IEEE.

[Sajadi et al., 2009] Sajadi, B., Lazarov, M., Gopi, M., and Majumder, A. (2009). Color seamlessness in multi-projector displays using constrained gamut morphing. *Visualization and Computer Graphics, IEEE Transactions on*, 15(6):1317–1326.

[Sanders and McCormick, 1987] Sanders, M. S. and McCormick, E. J. (1987). *Human factors in engineering and design* . McGRAW-HILL book company.

[Smith et al., 2014] Smith, J., Booth, T., and Bailey, R. (2014). Refresh rate modulation for perceptually optimized computer graphics.

[Staadt et al., 2006] Staadt, O. G., Ahlborn, B. A., Kreylos, O., and Hamann, B. (2006). A foveal inset for large display environments. In *Proceedings of the 2006 ACM international conference on Virtual reality continuum and its applications*, pages 281–288. ACM.

[Stockman et al., 1993] Stockman, A., MacLeod, D. I., and Johnson, N. E. (1993). Spectral sensitivities of the human cones. *JOSA A*, 10(12):2491–2521.

[Strasburger et al., 2011] Strasburger, H., Rentschler, I., and Jüttner, M. (2011). Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13.

[Stromeyer and Klein, 1974] Stromeyer, C. d. and Klein, S. (1974). Spatial frequency channels in human vision as asymmetric (edge) mechanisms. *Vision Research*, 14(12):1409–1420.

[Sundstedt et al., 2004] Sundstedt, V., Chalmers, A., and Cater, K. (2004). Selective rendering of task related scenes. In *Proceedings of the 1st Symposium on Applied perception in graphics and visualization*, pages 174–174. ACM.

[Sutherland, 1965] Sutherland, I. E. (1965). The ultimate display. *Multimedia: From Wagner to virtual reality*.

[Świrski et al., 2012] Świrski, L., Bulling, A., and Dodgson, N. (2012). Robust real-time pupil tracking in highly off-axis images. In *Proceedings of the Symposium on Eye Tracking Research and Applications*, pages 173–176. ACM.

[Tan and Czerwinski, 2003] Tan, D. S. and Czerwinski, M. (2003). Effects of visual separation and physical discontinuities when distributing information across multiple displays. In *Proc. Interact*, volume 3, pages 252–255.

[Tatler et al., 2005] Tatler, B. W., Baddeley, R. J., and Gilchrist, I. D. (2005). Visual correlates of fixation selection: Effects of scale and time. *Vision research*, 45(5):643–659.

[Tatler et al., 2011] Tatler, B. W., Hayhoe, M. M., Land, M. F., and Ballard, D. H. (2011). Eye guidance in natural vision: Reinterpreting salience. *Journal of vision*, 11(5):5.

[Tong and Fisher, 1984] Tong, H. and Fisher, R. (1984). Progress report on an eye-slaved area-of-interest visual display. Technical report, DTIC Document.

[Tsai and Liu, 2014] Tsai, W.-J. and Liu, Y.-S. (2014). Foveation-based image quality assessment. In *Visual Communications and Image Processing Conference, 2014 IEEE*, pages 25–28. IEEE.

[Tyler and Hamer, 1990] Tyler, C. W. and Hamer, R. D. (1990). Analysis of visual modulation sensitivity. iv. validity of the ferry–porter law. *JOSA A*, 7(4):743–758.

[Virsu and Rovamo, 1979] Virsu, V. and Rovamo, J. (1979). Visual resolution, contrast sensitivity, and the cortical magnification factor. *Experimental Brain Research*, 37(3):475–494.

[Wang et al., 2001] Wang, Z., Bovik, A. C., Lu, L., and Kouloheris, J. L. (2001). Foveated wavelet image quality index. In *International Symposium on Optical Science and Technology*, pages 42–52. International Society for Optics and Photonics.

[Wang et al., 2004] Wang, Z., Bovik, A. C., Sheikh, H. R., and Simoncelli, E. P. (2004). Image quality assessment: from error visibility to structural similarity. *Image Processing, IEEE Transactions on*, 13(4):600–612.

[Watson et al., 1997] Watson, B., Walker, N., Hodges, L. F., and Worden, A. (1997). Managing level of detail through peripheral degradation: Effects on search performance with a head-mounted display. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 4(4):323–346.

[Wilson, 1980] Wilson, H. R. (1980). A transducer function for threshold and suprathreshold human vision. *Biological cybernetics*, 38(3):171–178.

[Yang et al., 2001] Yang, R., Gotz, D., Hensley, J., Towles, H., and Brown, M. S. (2001). Pixelflex: A reconfigurable multi-projector display system. In *Proceedings of the conference on Visualization'01*, pages 167–174. IEEE Computer Society.